# Representing Knowledge for Logic-based Diagnosis

**David Poole**
Department of Computer Science,
University of British Columbia,
Vancouver, B. C., Canada, V6T1W5
poole@cs.ubc.ca

## Abstract

If one wants to use logic to build a diagnostic system, then it is not a matter of "just axiomatising" the domain; we have to understand how to use logic for diagnosis. We need some models of what diagnosis is, in order to be able to implement diagnostic systems. This paper considers 3 different "logical" definitions of diagnosis. Each of these are presented in a uniform framework of hypothetical reasoning where the user provides the possible hypotheses. These are compared as to the sort of knowledge that we need to provide them, and in their expressibilty. It seems as though there is no one framework which can claim to be the logical definition of diagnosis.

Each of these approaches has been implemented in the Theorist system, and used on a number of domains. This paper concentrates on the case where we have fault models.

## 1 Introduction

Diagnosis is a problem of trying to find what is wrong with some system based on knowledge about the design/structure of the system, possible malfunctions that can occur in the system and observations (symptoms, evidence) made of the behaviour of the system.

There seems to be three predominant approaches to the problem of diagnosis:

1. minimising assumptions of abnormal components that are consistent with all knowledge and observations [Reiter87, de Kleer87, Davis84, Genesereth84].
2. abductive diagnosis, finding a set of causes which can imply the observations [PGA87, Cox87, Reggia83].
3. rule-based diagnosis, where we have a set of symptom–cause rules, and want to determine what malfunctions we can predict based on the evidence [Buchanan84, Pearl87a].

These seem to be few comparisons of these discussing how they can be used to perform diagnosis, what assumptions about the sorts of knowledge they each use, in what ways they are similar and different. Other comparisons of diagnostic procedures (eg., [Ramsey85, Koton85]) have been more concerned with informal analysis of how they worked on a few examples, rather than comparing underlying assumptions. This paper is an attempt to fill this void.

As a point of terminology, I will refer to an *approach* to diagnosis as an abstract idea behind a form of diagnosis (eg., the idea of abductive diagnosis); a *theory* of diagnosis as a specification of the formal definition of diagnosis (eg, [Reggia83], [Cox87] and [PGA87] each specify different theories of abduction); and a *system* as any implementation of a theory. We would like to talk of properties of all implementations of a particular theory; once we know what is the correct specification of diagnosis is (or at least what the tradeoffs are) then we can concentrate on computing it efficiently.

## 2 The Theorist Framework

Theorist [Poole88, PGA87] is a theory and implementation of default and abductive reasoning. It is based on a restricted form of hypothetical reasoning, namely where the user provides the system with a set possible hypotheses they are prepared to accept in an explanation as to why something may be true.

This formalism is suited to the task of understanding diagnostic tasks as it allows for default and abductive reasoning in a uniform, formal framework. Each of the three sorts of reasoning is easily expressible in the Theorist framework. Note that a commitment to the Theorist framework is not a commitment to any particular control structure (search strategy).

The Theorist system is provided with two sets of first order formulae:

$F$ is a set of closed formulae called the *facts*. These are intended to be true in the world being modelled.

$H$ is a set of formulae which act as *possible hypotheses*, any ground instance of which can be hypothesised if consistent.

**Definition 2.1** a **scenario** of $A, H$ is a set $D \cup A$ where $D$ is a set of ground instances of elements of $H$ such that $D \cup A$ is consistent.

That is, a scenario is any consistent set of assumptions.

**Definition 2.2** If $g$ is a closed formula then an **explanation** of $g$ from $A, H$ is a scenario of $A, H$ which implies $g$.

That is, $g$ is explainable from $A, H$ if there is a set $D$ of ground instances of elements of $H$ such that

$A \cup D \models g$ and
$A \cup D$ is consistent

$A \cup D$ is an explanation of $g$.

**Definition 2.3** an **extension** of $A, H$ is the set of logical consequences of a maximal (with respect to set inclusion) scenario of $A, H$.

In [Poole88] the correspondence between this definition of extensions and the definition of [Reiter80] (where $\delta \in H$ corresponds to the default $: \delta/\delta$ in [Reiter80]) is proved.

## 3 The Diagnostic Models

Any diagnosis system requires knowledge about the domain of diagnosis and observations of the actual artifact we are diagnosing. The sort of knowledge that is required can be divided into:

**domain model** which describes the structure of the system, how normal components work, how abnormal components work, and how faults manifest themselves. In all of the systems these will correspond to statements which are constrained to be true of the artifact being diagnosed (i.e., we have enough caveats to make them facts, eg., by saying "if this component is normal and this other component has such and such a fault which is acting normally for that fault then ... ").

**observations** is the set of observations made of the actual artifact we are diagnosing.

**normality assumptions** are hypotheses that some component is working correctly.

**abnormality assumptions** are hypotheses that some component is not working correctly. This can be seen as the negation of a normality assumption.

**fault assumptions** are assumptions of some particular fault or disease. There may be many different faults possible for an abnormality; and one fault may imply many components are abnormal. A fault can often be seen as a cause for why some components are acting abnormally.

We first begin with our three definitions of diagnosis, together with their translation into the Theorist framework.

**Definition 3.1** A **diagnosis**$_1$ is minimal set of abnormality assumptions such that the observations are consistent with all other components acting normally[Reiter87].

In terms of the Theorist framework,

$F$ is the domain model together with the observations.

$H$ is the set of normality assumptions.

**a diagnosis** corresponds to an *extension* (in particular, the set of abnormality assumptions in an extension) [Reiter87, theorem 6.1].

**Definition 3.2** A **diagnosis**$_2$ is a minimal set of assumptions which implies the observations [PGA87].

In terms of the Theorist framework,

$F$ is the domain model.

$H$ is the set of normality assumptions and fault assumptions.

**a diagnosis** is an *explanation* of the observations.

**Definition 3.3** A **diagnosis**$_3$ is a set of fault conditions (possible malfunctions) which can be explained from structure and the observations.

In terms of the Theorist framework,

$F$ is the domain model together with the observations.

$H$ is a set *symptom*$\Rightarrow$*cause* rules. By being part of the possible hypotheses, these act as defaults [Poole88].

**a diagnosis** is the set of fault assumptions that can be explained.

## 4 Using the Diagnosis Systems

### 4.1 What sort of knowledge is required?

Before we can do any detailed comparisons of the diagnostic theories we need to consider how one would go about applying each diagnostic system to solving problems.

There seems to be two extremes as to the sort of knowledge that one may have of a domain:

1. We have knowledge about how components are structured and work normally. There is no knowledge as to how malfunctions occur and manifest themselves. The system is described totally in terms of normality conditions.
2. We have just information on faults (diseases) and their symptoms, and want to account for the abnormal observations [Reggia83].

It is instructive to examine how both normality condition and fault models can be used by each of the diagnostic systems. In this paper we concentrate on how fault models can be used by each of the diagnostic theories.

### 4.2 Causes and Symptoms

As part of the terminology for talking about the domain, I will use the terms "causes" and "symptoms". Causes can be seen as reasons why the symptom occurred. In this paper we are not assuming any theory of causality; a theory of causality is imposed by the builder of the knowledge base (the person who models the system being diagnosed). We want to allow as much flexibility as possible in the interpretation of these terms.

Note that the terms "cause" and "symptom" are internal and local terms. It is quite conceivable (and indeed very common) that something is seen as both a cause for some symptom, and something which needs to be explained as a symptom. For example, we may see someone coughing (a symptom) and have as a cause, that the person has a sore throat. We may then have a viral infection as the cause for the symptom of sore throat.

I will use the terms of "base cause" for the causes which don't need any further explanation (it is up to the user to determine what these are), and "observed symptom" (or just "observation" for the symptom that we actually have observed.

### 4.3 Fault Models

Diagnosis1 is defined in terms of normality assumptions rather than in terms of fault models. The other two diagnostic models are in terms of fault models. Before we can offer a detailed comparison, we have to consider how we could incorporate fault models into diagnosis1.[1]

To add fault models to diagnosis1, there is the question of what should be minimised (its negation assumed) and maximised (assumed). There seems to be two alternatives

1. to maximise normality and minimise abnormality and to let fault assumptions be minimised as a side effect of minimising abnormality. Faults in this model are just incidental to the diagnosis, and can only be used to rule out abnormalities as there may be no cause for that abnormality.
2. to assume the negation of a fault assumption as a possible hypothesis. This is, in fact what is done in [Reiter87] to model the generalised set covering model of [Reggia83]. In this paper I assume that this is the approach taken.

It is important to note that the diagnoses are the faults that can be proven from the assumption that other faults are absent [Reiter87, proposition 3.3].

### 4.4 Representing Causes

First let us examine how we can represent and reason about fault models in each of the systems. Fault models are closely related to finding what is causing the problems being manifested.

We first want to consider the question *what sort of knowledge is required?* We consider each of the diagnostic theories in turn.

1. In diagnosis1, we have to prove an abnormality (maybe based on other assumptions). Thus the sort of knowledge we need is of the form $obs \Rightarrow ab$ (or $\neg ab \Rightarrow predn$).

   Knowledge of the form $ab \Rightarrow symptoms$ cannot be used to conclude (or hypothesise) some abnormality, it can only be used to rule out a possible cause.

   In terms of faults we have to specify conditions to be met before we can conclude a fault (as we have to end up proving a fault from the assumption of the absence of other faults). The possible hypotheses are the negations of the base causes.
2. In diagnosis2, the sort of knowledge we need is that from some explanation we can prove the observations. Thus the sort of knowledge is of the form $fault \Rightarrow symptoms$. The base faults become the possible hypotheses.
3. In diagnosis3, we have to explain a fault. Thus the sort of knowledge is of the form $obs \Rightarrow ab$, usually with default status (i.e., it is a possible hypothesis).

If $c_1, ..., c_n$ are the possible causes we are prepared to accept as an explanation of why symptom $s$ occurred then for each of the systems we give knowledge

1. For diagnosis1 we have as a fact or a default $s \Rightarrow c_1 \vee ... \vee c_n$. That is, if we have symptom $s$ then it is inconsistent that they do not have any of the $c_i$.
2. In diagnosis2, the sort of knowledge we need is stating that from some explanation we can prove the observations. Thus the sort of knowledge is of the form $c_i \Rightarrow s$ (this can either be a fact or a possible hypothesis).
3. For diagnosis3, we represent $s \Rightarrow c_i$ as a default. If we observe $s$ then this, by default, is evidence for $c_i$.

**Example 4.1** Consider representing the following knowledge about how aching elbows and aching hands could be caused:

> *tennis-elbow* causes *aching-elbow*
> *dishpan-hands* causes *aching-hands*
> *arthritis* causes both *aching-elbow* and *aching-hands*

[1] It should be emphasised here that what I mean as an abnormality is a statement that some component is not working correctly. One reading of Reiter's paper is that an abnormality is whatever we are minimising. I use a more precise definition.

Consider how such knowledge can be expressed so that it can be used by each of the diagnostic systems:

1. For diagnosis1, we can represent the above knowledge by having

$$\begin{aligned} H &= \{ \neg \textit{tennis-elbow}, \neg \textit{dishpan-hands}, \neg \textit{arthritis}\} \\ F &= \{ \textit{aching-elbow} \Rightarrow \textit{tennis-elbow} \vee \textit{arthritis}, \\ &\quad\ \ \textit{aching-hands} \Rightarrow \textit{dishpan-hands} \vee \textit{arthritis}\} \end{aligned}$$

   If we observe aching-elbow then it must have been caused by either tennis-elbow or by arthritis.

2. For diagnosis2, we have

$$\begin{aligned} H &= \{ \textit{tennis-elbow, dishpan-hands, arthritis}\} \\ F &= \{ \textit{tennis-elbow} \Rightarrow \textit{aching-elbow}, \\ &\quad\ \ \textit{dishpan-hands} \Rightarrow \textit{aching-hands}, \\ &\quad\ \ \textit{arthritis} \Rightarrow \textit{aching-elbow} \wedge \textit{aching-hands}\} \end{aligned}$$

   Thus we are representing the causal knowledge as implications.

3. For diagnosis3, we have the following evidential rules:

$$\begin{aligned} H &= \{ \textit{aching-elbow} \Rightarrow \textit{tennis-elbow}, \\ &\quad\ \ \textit{aching-hands} \Rightarrow \textit{dishpan-hands}, \\ &\quad\ \ \textit{aching-elbow} \Rightarrow \textit{arthritis}, \\ &\quad\ \ \textit{aching-hands} \Rightarrow \textit{arthritis}\} \end{aligned}$$

   (or, perhaps, the last one should be *aching-elbow* $\wedge$ *aching-hands* $\Rightarrow$ *arthritis*). Thus *tennis-elbow* causes *aching-elbow* and so *aching-elbow* is, by default (i.e., unless there are other reasons for ruling it out) evidence for *tennis-elbow*.

Suppose we observe *aching-elbow*; consider what we conclude from each of the diagnosis systems:

1. There are two extensions, one containing

$$\{\neg \textit{tennis-elbow}, \neg \textit{dishpan-hands}, \textit{arthritis}\}$$

   and one containing

$$\{\textit{tennis-elbow}, \neg \textit{dishpan-hands}, \neg \textit{arthritis}\}$$

2. to explain *aching-elbow* we have two explanations:

$$\{\textit{tennis-elbow}\}$$
$$\{\textit{arthritis}\}$$

3. We can explain *tennis-elbow*, and *arthritis*. Here there is one extension, containing

$$\{\textit{tennis-elbow, arthritis}\}$$

Consider observing *aching-elbow*$\wedge$*aching-hands*. In this case we conclude from each of the diagnosis systems:

1. There are two extensions, one containing

$$\{\neg \textit{tennis-elbow}, \neg \textit{dishpan-hands}, \textit{arthritis}\}$$

   and one containing

$$\{\textit{tennis-elbow, dishpan-hands}, \neg \textit{arthritis}\}$$

2. to explain *aching-hands*$\wedge$*aching-elbow* we have two explanations:

$$\{\textit{tennis-elbow, dishpan-hands}\ \}$$
$$\{\textit{arthritis}\}$$

3. We can explain *tennis-elbow, dishpan-hands* and *arthritis*. Here there is one extension, containing

$$\{\textit{tennis-elbow, dishpan-hands, arthritis}\}$$

This example can be very instructive on the differences between the diagnostic systems. The extensions of diagnosis1 and the explanations of diagnosis2 seem to be very similar (in section 4.6 this equivalence is spelled out in greater detail). Diagnosis3 seems to be the odd one out; in diagnosis3 we lost the structure of the evidence; this turns out to be a general trend.

### 4.5 Ruling out Causes

What sort of knowledge do we need to rule out particular causes from consideration? For example ruling out sulphuric acid as a pollutant of a stream because there is no sulphates in the water samples.

To have this sort of knowledge in any of the systems we need to have knowledge of the form

$$\textit{evidence} \Rightarrow \neg \textit{cause}$$

These are "causal rules" because they give the implication of the symptoms from the causes. This is the sort of knowledge that diagnosis2 needed in the first place, but is the opposite sort of implication than I claimed before that was needed in diagnosis1 or diagnosis3. Thus it seems as though in a system for diagnosis1 or diagnosis3 one needs both causal rules and evidential rules.

Thus if $c_1, ..., c_n$ are the possible causes of $s$, then diagnosis2 needs knowledge of the form

$$c_1 \Rightarrow s, ..., c_n \Rightarrow s$$

whereas diagnosis1 needs that knowledge as well as knowledge of the form

$$s \Rightarrow c_1 \vee ... \vee c_n$$

Of course, there is much more subtlety in the sort of knowledge used by each system. It is however instructive to consider an idealised "standard" case, and then to consider how each of them can deviate from the standard case.

### 4.6 Standard Propositional case

The standard case we will consider first is where all of the knowledge is propositional and the symptoms of the diseases are definite (i.e., a cause always causes some symptom), and we have complete knowledge. From understanding this simple case, we can then learn about more complicated cases.

Suppose that for possible symptom $s$, we have causes $c_1, ..., c_n$ (each of these can be a conjunction of base causes or even other non-base causes, which themselves have to be explained).

As discussed above, the sort of knowledge that we need for diagnosis1 is of the form $s \Rightarrow c_1 \vee ... \vee c_n$ in order to conclude a cause, together with $c_i \Rightarrow s$ for each $i$ in order to rule out possible causes. Thus it is of the form

$$s \equiv c_1 \vee ... \vee c_n$$

The sort of knowledge we need for diagnosis2 is of the form

$$(c_1 \Rightarrow s) \wedge ... \wedge (c_n \Rightarrow s)$$

Notice that the first looks just like the completion (in terms of [Clark78]) of the second. It will turn out to be closely related, but there are two important differences

1. If $c$ is a basic cause, then we don't want to complete it. There may not be any formulae which imply $c$, but we do not want to then say that $c$ is false (as we would in the full completion).
2. We are not only working with what [Lloyd87] calls "program statements"; we want to be able to say that someone does not have some symptom, this can then be used to prune our set of explanations. We thus have explicit negation and not just negation as failure.

If we have $F$ as the facts and $H$ as the possible hypotheses for diagnosis2, then define the **completion** of $F$ with respect to $H$ to be the $F$ together with, for each $a$ which is not an element of $H$, the formulae $a \Rightarrow c_1 \vee ... \vee c_n$, where $(c_1 \Rightarrow a) \wedge ... \wedge (c_n \Rightarrow a)$ is the set of formulae in $F$ which imply $a$.

Each of the diagnostic system can however express more subtlety than the form given above. For diagnosis1, we do not have to state the logical equivalence between the symptom and the disjunct of possible causes For example, we may say that some cause could possibly have caused a symptom, but the symptom is not a necessary part of that cause. Without this sort of knowledge we can never prune the set of symptoms based on missing symptoms. This can be expressed in diagnosis2 by making the implication $c \Rightarrow s$ as a possible hypothesis which can be hypothesised to explain $s$ (but is not used to reject $c$ if we can show $\neg s$).

Thus if $c_1, ..., c_n$ are the possible causes of symptom $s$, then diagnosis1 would represent this as $s \Rightarrow c_1 \vee ... \vee c_n$, and for each $c_i$ for which $s$ is a necessary symptom, we have $c_i \Rightarrow s$ as a fact. Diagnosis2 would represent this as $c_i \Rightarrow s$ being a fact if $s$ is a necessary symptom of $c_i$, and $c_i \Rightarrow s$ as a possible hypothesis otherwise. Any other relationship between the two (eg., a cause implying a disjunct of symptoms) would be added as facts to each of these.

Under these conditions it turns out that the diagnoses are identical. We assume that the knowledge bases are in their simplest form, where there are no causality loops. This seems like a reasonable assumption for cases where we are axiomatising causality.

**Theorem 4.2** *If $K1$ is the knowledge base for diagnosis1, and $K2$ is the corresponding knowledge base for diagnosis2, then the diagnoses using diagnosis2 from $K2$ are identical to the diagnoses using diagnosis1 from $K1$.*

**Proof:** This is proven by induction on the number of atomic symbols in the knowledge base.

If there is only one atomic symbol, $a$, that is observed, then there are two cases to consider

1. it is a basic cause. In this case, if its negation is provable from $K2$ then there are no diagnoses in either case. Otherwise, in both cases there is the diagnosis $\{a\}$.
2. it is not a basic cause. In this case, if it is provable from $K2$, then we have the empty diagnosis for each system. If it is not provable from $K2$, then there is no diagnosis for diagnosis $K2$, and in $K1$, there must be the fact $a \Rightarrow false$ (as there is nothing to prove $a$), so $K1$ is inconsistent with the observation $a$, so there again is no diagnosis.

Suppose that $s_1, ..., s_n$ are our symptoms to be explained. If $n = 0$, the empty diagnosis is a diagnosis for each system. If $s_1$ is not a base cause, there will be a (possibly empty) set of rules $c_i \Rightarrow s_1$ in $K2$. Now consider $K2'$ which is $K2$ with these rules removed, and $K1'$ as $K1$ with the corresponding rules and the completion rule removed. Consider the explanations of the symptoms $c_i, s_2, ..., s_n$ for each $i$. We have thus created a system with one less atomic symbol (we have removed all rules about $s_1$). By the inductive assumption, the diagnoses from $K1'$ and $K2'$ are identical. Suppose, that for each $i$, these are $D^i_1, ..., D^i_{k^i}$. The diagnoses of $s_1, ..., s_n$ from $K2$ consist of the subset of these that are consistent (as each diagnosis must prove all of the goals). These could only be inconsistent by those rules of the form $c_i \Rightarrow s_1$ that are facts. These are also inconsistent with $K1$ (as $K2 \subseteq K1$), and so are not diagnoses using diagnosis1. In $K1$ is the rule $s_1 \Rightarrow c_1 \vee ... \vee c_m$, and so we have $s_1 \Rightarrow \bigvee_{\{i,j\}} D^i_j$, and so $s_1$ implies the disjunct of all of those that are consistent, and so each minimal $D^i_j$ that is consistent is a diagnosis (as we can prove that, from the assumption that all other causes are absent, that diagnosis).

□

Differences still arises if the sort of knowledge is not of the form of our standard case. It is important to note how the standard case works when there is no possible causes of a symptom. In the analysis above, for diagnosis2, this means that we cannot explain the symptom; for the representation for diagnosis1 we have stated that the symptom could not occur (it implies the empty disjunction, which is false).

Differences still arise, for example if the knowledge base contains $ab\ a \vee ab\ b$, and there are no observations. In diagnosis2, if there are no observations, then there is always the empty diagnosis if the knowledge base is consistent. For diagnosis1, there is no distinction between the general knowledge and the observations, and so there is nothing special about the relationship between the observations of the artifact being diagnosed and the diagnoses. In the case with $ab\ a \vee ab\ b$ as the knowledge base, there are two diagnoses ($\{ab\ a\}$ and $\{ab\ b\}$), even with no observations. Why and how one may want to exploit such distinctions is still an open question.

### 4.7 Pearl's example

**Example 4.3 (Pearl)** Pearl [Pearl87a, p. 371] gives the following example (in the context of diagnosis3) to argue that there should be a distinction between *causal rules* and *evidential rules*. Here we show how the problems he was trying to solve in diagnosis3 do not arise in diagnosis1 and diagnosis2.

The knowledge we want to represent is of the form

*rained-last-night* causes *grass-is-wet.*
*sprinkler-was-on* causes *grass-is-wet.*
*grass-is-wet* causes *grass-is-cold-and-shiny.*
*grass-is-wet* causes *shoes-are-wet.*

Each of the diagnosis systems would represent this knowledge as

1. For diagnosis1, we would represent this as

$$\begin{array}{rll} F & = \{ & \textit{grass-is-wet} \equiv \textit{sprinkler-was-on} \vee \textit{rained-last-night}, \\ & & \textit{grass-is-wet} \equiv \textit{grass-is-cold-and-shiny}, \\ & & \textit{grass-is-wet} \equiv \textit{shoes-are-wet}\} \\ H & = \{ & \neg\textit{rained-last-night}, \neg\textit{sprinkler-was-on}\} \end{array}$$

2. For diagnosis2, we would represent the same knowledge as

$$\begin{array}{rll} F & = \{ & \textit{rained-last-night} \Rightarrow \textit{grass-is-wet}, \\ & & \textit{sprinkler-was-on} \Rightarrow \textit{grass-is-wet}, \\ & & \textit{grass-is-wet} \Rightarrow \textit{grass-is-cold-and-shiny} \wedge \textit{shoes-are-wet}\} \\ H & = \{ & \textit{rained-last-night}, \textit{sprinkler-was-on}\} \end{array}$$

3. For diagnosis3, we would represent the same knowledge as

$$\begin{array}{rll} H & = \{ & \textit{rained-last-night} \Rightarrow \textit{grass-is-wet}, \\ & & \textit{sprinkler-was-on} \Rightarrow \textit{grass-is-wet}, \\ & & \textit{grass-is-wet} \Rightarrow \textit{sprinkler-was-on}, \\ & & \textit{grass-is-wet} \Rightarrow \textit{rained-last-night}, \\ & & \textit{grass-is-wet} \Rightarrow \textit{grass-is-cold-and-shiny}, \\ & & \textit{grass-is-cold-and-shiny} \Rightarrow \textit{grass-is-wet}, \\ & & \textit{grass-is-wet} \Rightarrow \textit{shoes-are-wet}, \\ & & \textit{shoes-are-wet} \Rightarrow \textit{grass-is-wet}\} \end{array}$$

Suppose that we observe that it rained last night, then for each of the systems we get

1. there is one extension containing

   $$\{\textit{rained-last-night}, \neg\textit{sprinkler-was-on}\}$$

   From this we can prove that the grass is wet, that the grass is cold and shiny and that my shoes are wet.

2. there is one explanation of *rained-last-night*, namely

   $$\{\textit{rained-last-night}\}$$

   From this we can prove that the grass is wet, that the grass is cold and shiny and that my shoes are wet.

3. we can explain everything, including that the sprinkler was on last night. [Pearl87a] attributes this problem to not distinguishing between evidential and causal rules. I would claim that it is a flaw in the idea of diagnosis3.

If we had instead observed that the grass is cold and shiny, then we get:

1. there are two extensions,

   $$\{\textit{rained-last-night}, \neg\textit{sprinkler-was-on}\}$$

   $$\{\neg\textit{rained-last-night}, \textit{sprinkler-was-on}\}$$

2. there are two explanations

   $$\{\textit{rained-last-night}\}$$

   $$\{\textit{sprinkler-was-on}\}$$

3. we can explain both *rained-last-night* and *sprinkler-was-on.*

From all of these we can predict that my shoes are wet.

## 5 Uncertainty

The analysis we have considered about diagnostic theories is orthogonal to the problem as to what is a "better" diagnosis.

All three of these diagnostic systems have been imbued with uncertainty calculus. In particular each of them has had a probability measure associated with them. For example [de Kleer87] associates a conditional probability with a candidate; [Neufeld87] associates a conditional probability with an explanation;

the evidential rules of [Pearl87a] can be seen as being derived from conditional probabilities [Pearl87b].

The interesting thing about this is that none of the methods have a special claim to be the approach sanctioned by probability. Each of them specifies a different set of formulae we want to get the probability of.

## 6 Conclusion

In this paper I have examined three different ways to think about diagnosis. It seems as though there is no right or wrong definition of diagnosis. Which is better depend on which one thinks contains a more natural representation of the systems being diagnosed.

It was shown that for the propositional case using fault models, that two of the diagnostic systems were essentially equivalent. A few differences were:

1. In diagnosis1 we have to explicitly make the complete knowledge assumption; we could not use the system if we did not enumerate the list of possible causes. For diagnosis2, we did not need to make any such assumption. If we wanted to interpret the set of diagnoses as covering then we needed to have a complete knowledge assumption, but there was nothing in the formalism nor in the way that it is used that forces us to interpret the set of diagnoses as covering.
2. The sort of knowledge for diagnosis2 is much more modular than that for diagnosis1. It seems as though we are more likely to have information of the symptoms of diseases than have knowledge of what are all of the possible causes of some symptom. Diagnosis1 requires all of the knowledge initially, and adding new knowledge requires debugging of the knowledge base, rather than just the modular addition of knowledge.
3. diagnosis1 requires us to make assumptions that are irrelevant to the observations, for example, when we observed aching elbow in example 4.1, the the diagnosis assumed that we did not have an aching elbow. This can be fixed up, by considering the diagnoses as the generators of all supersets of the diagnoses (as in [de Kleer87]), but then the definition seems to be different to that given in [Reiter87].
4. One of the requirements of a logical definition of diagnosis, is that we do not want to have to write as facts things which are not true of the intended interpretation. In this respect, diagnosis2 fares much better than diagnosis1. In diagnosis1 we have to make the complete knowledge assumption in writing down what was true about the domain, as opposed to making the complete knowledge assumption only in how the diagnoses are interpreted (as in diagnosis2).

From the analysis in example 4.1 and example 4.3, it seems as though there is something wrong with diagnosis3. It loses the structure in the problem, and does not allow a natural interpretation of the results. Work like [Pearl87a] may fix up the problems, but it is not clear that it is worth patching up.

This paper is not intended to be a definitive comparison of the diagnostic paradigms. There are a number of cases which still need to be considered, including the case with variables, the case where we observe a system with inputs as well as outputs, the problem of discriminating between diagnoses, and empirical results as to how they each perform in practice. More work needs to be done, and more work is under way.

## Acknowledgements

This work was carried out while the author was at the University of Waterloo. Thanks to the members of the LPAIG for many discussions and arguments about Theorist, especially to Randy Goebel, Eric Neufeld, Paul Van Arragon and Romas Aleliunas. I am also indebted to Ray Reiter and Johan de Kleer for providing clear accounts of the logic of their systems, and to Matt Ginsberg for forcing me to consider how the two forms of diagnosis really do differ. This work was supported under NSERC grant A6260.

## References

[Buchanan84] B. G. Buchanan and E. H. Shortliffe, *Rule-Based Expert Systems*, Addison-Wesley, Reading, MA.

[Clark78] K.L.Clark, "Negation as Failure", in H.Gallaire and J.Minker (Eds), *Logic and Databases*, Plenum Press, New York, 293-322.

[Cox87] P. T. Cox and T. Pietrzykowski, *General Diagnosis by Abductive Inference*, Technical report CS8701, School of Computer Science, Technical University of Nova Scotia, April.

[Davis84] R. Davis, "Diagnostic Reasoning Based on Structure and Behaviour", *Artificial Intelligence 24*, pp. 347-410.

[de Kleer87] J. de Kleer and B. C. Williams, "Diagnosing Multiple Faults", *Artificial Intelligence, Vol. 32, No. 1*, pp. 97-130.

[Genesereth84] M. R. Genesereth, "The Use of Design Descriptions in Automated Diagnosis", *Artificial Intelligence*, Vol. 24, pp. 411-436.

[Koton85] P. A. Koton, "Empirical and Model-based reasoning in expert systems", *Proc. IJCAI85*, pp. 297-299.

[Lloyd87] J. W. Lloyd, *Foundations of Logic Programming*, Second Edition, Springer-Verlag.

[Neufeld87] E. M. Neufeld and D. Poole, "Towards solving the multiple extension problem: combining defaults and probabilities", *Workshop on Reasoning with Uncertainty*, Seattle, July.

[Pearl87a] J. Pearl, "Embracing Causality in Formal Reasoning", *Proc. AAAI-87*, pp. 369-373.

[Pearl87b] J. Pearl, "Distributed Revision of Composite Beliefs", *Artificial Intelligence*, Vol. 33, No. 2.

[PGA87] D. L. Poole, R. G. Goebel, and R. Aleliunas, "Theorist: a logical reasoning system for defaults and diagnosis", in N. Cercone and G.McCalla (Eds.) *The Knowledge Frontier: Essays in the Representation of Knowledge*, Springer Varlag, New York, pp. 331-352.

[Poole88] D. L. Poole, "A Logical Framework for Default Reasoning", *Artificial Intelligence*, Vol. 36, No. 1, pp. 27-47, August 1988.

[Ramsey85] C. L. Ramsey, J. A. Reggia, D. S. Nau and A. Ferrentino, *A comparative analysis of methods for expert systems*, Technical Report, University of Maryland, TR-1491.

[Reiter80] R. Reiter, "A Logic for Default Reasoning", *Artificial Intelligence*, Vol. 13, pp 81-132.

[Reiter87] R. Reiter, "A Theory of Diagnosis from First Principles" *Artificial Intelligence*, Vol. 32, No. 1, pp. 57-96.

[Reggia83] J. A. Reggia, D. S. Nau and P. Y. Wang, "Diagnostic expert systems based on a set covering model", *Int. J. Man-Machine Studies 19*, pp.437-460.