# Representing Diagnosis Knowledge

David Poole
Department of Computer Science,
University of British Columbia,
Vancouver, B. C., Canada, V6T 1Z2
and Canadian Institute for Advanced Research
poole@cs.ubc.ca

March 2, 1993

**Abstract**

This paper considers the *representation problem*: namely how to go from an abstract problem to a formal representation of the problem. We consider this for two conceptions of logic-based diagnosis, namely abductive and consistency-based diagnosis. We show how to represent diagnostic problems that can be conceptualised causally in each of the frameworks, and show that both representations of the same problems give the same answers. This is a local transformation that allows for an expressive (albeit propositional) language for giving the constraints on what symptoms and causes can coexist, including non-strict causation. This non-strict causation can be represented in each framework *without* adding special reasoning constructs to either framework. This is presented as a starting point for a study of the representation problem in diagnosis, rather than as an end in itself.

## 1 Introduction

This paper defines an abstract "knowledge representation" problem and considers the problem of representing knowledge in the context of diagnostic systems. We consider two diagnostic formalisms and compare how we can

represent a domain in each so that each representation produces the same answer.

This paper contains many of the results of [14], recast in light of latter developments. We take a different perspective from subsequent (to [14]) papers [2, 3, 9], in that we consider the problem of going from an abstract problem to a representation of the problem rather than the problem of just going from one representation to another. While the local transformation methods may work for simple theories, there is still much to be learnt about what needs to go into any axiomatisation [16], and the mappings are not so straight forward.

One of the ideas that we are tackling is to represent subtle distinctions in the domain with rather weak representation languages. One of the main reasons for pushing weak representation languages is that we can see what they can and cannot represent, and only complicate the representations when necessary. In this paper we consider how to represent causal relations that are not strict implications (e.g., a cold may cause sneezing, but it does not *imply* sneezing). There are no new non-strict implications in either representation language we consider, but they can both represent strict and non-strict causes.

Like Console et. al., [2, 3] and unlike Konolige [9] we consider acyclic causal structures (some $c$ cannot cause itself). Acyclicity allows us to have a local transformation from the domain knowledge to the representations unlike the global transformations of Konolige (see [9, section 5.2]).

This paper does not contain the final answer to this problem; there is still much that has to be understood about representing more complex problems than that considered here [16].

## 1.1 The Knowledge Representation Problem

**Definition 1.1** Given a formalism (formal language plus an inference relation), the **knowledge representation problem** is the problem of going from a problem $P$ to a representation $R_P$ of $P$ in the formal language so that the use of the inference relation for the representation will yield a solution to the problem.

In this definition, a problem is "a question raised for inquiry, consideration, or solution" (definition from Webster's Ninth New Collegiate Dictio-

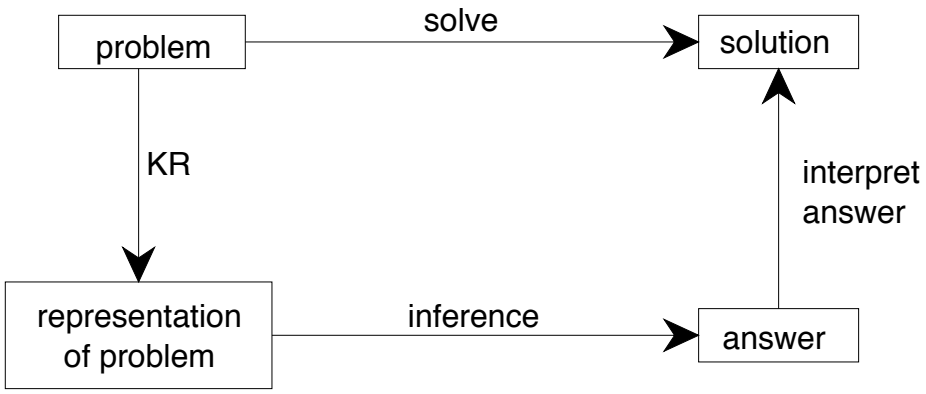


Figure 1: The knowledge representation problem.

nary). This is *not* a formal representation of the problem. Many problems can be conceptualized in different ways, and these different conceptualizations may have different representations (even for the same formalism) and may have very different computational and ergonomic properties.

Definition 1.1 is depicted in Figure 1. We want to define the knowledge representation (KR), the inference relation and the interpretation for the answers so that this diagram commutes[1].

This notion of knowledge representation should be contrasted with the view of knowledge representation (KR) research as defining and analysing formalisms, without the knowledge representation (as defined here) being explicit (see e.g., much of the work on nonmonotonic reasoning [8]). The knowledge representation problem is often implicit, defined in terms of a few examples of how to represent a particular problem[2].

As, by "the problem", we mean the problem itself and not a representation of the problem, it may seem that the knowledge representation problem

[1]A diagram commutes if each directed path to the same point produces that same answer. In this case, the solution to the problem obtained by going via the representation and computation is the same as the solution obtained going directly from the problem to the solution.

[2]I do not want to imply that I am defining a new KR problem; I am trying to be explicit about what I consider the KR problem to be. There are many instances of this view of KR from foundational papers (e.g., [12]) to textbooks based on this view (e.g., [5]).

cannot be formalised, or that there is nothing precise that can be said about the knowledge representation problem (until it itself is formalised and represented). I believe that that this view is mistaken. For example, one sort of things that can be said about the KR-problem is "if the problem can be conceptualized in some particular way, then it can be represented in some particular way". There may be many different representations of the same conceptualization, and many possible conceptualizations of the same problem. The different resulting representations can be compared in terms of efficiency (both computational efficiency and conceptual efficiency), and naturalness of the resulting representations.

This enterprise seems much more important when we realize that any logic that incorporates definite clauses, and for which logical consequence is allowed as part of the inference relation, is Turing equivalent and so can represent any problem (any computable problem can be encoded in definite clauses). For such (quite weak) representational formalisms, the question of representational adequacy [12] seems moot without explicitly considering the KR problem.

## 1.2 KR for diagnosis

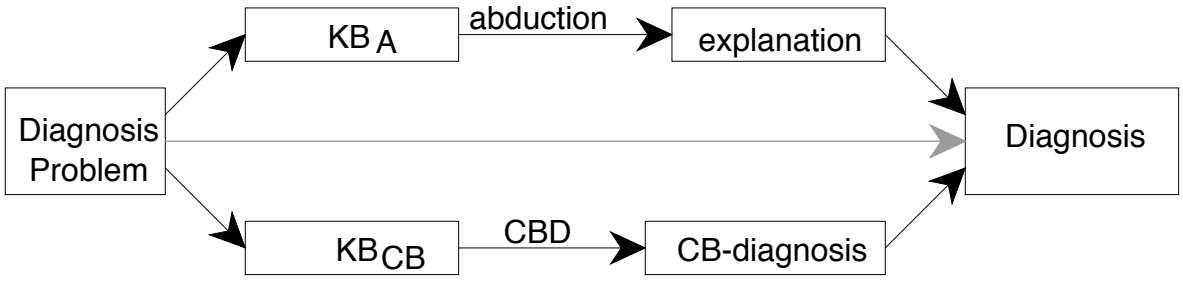


Figure 2: Diagnostic problem representations.

When considering KR for diagnosis, we consider two different formalisms (they have the same language, but have a different notion of what an answer is and thus need different "inference" mechanisms). This is shown in Figure 2. The two formalisms are (see Section 2 for formal definitions):

Abductive diagnosis — the answer is an explanation of the observations using abduction from an abductive KB ($KB_A$) (see Figure 2).

Consistency-based diagnosis (based on the model of Reiter [23]) — an answer is a consistency based diagnosis (CB-diagnosis in Figure 2), from a knowledge base ($KB_{CB}$ in Figure 2), using some way of computing CD-diagnoses (CBD in Figure 2)[3].

The main result of this paper is to show that, for a certain class of problems, the two representation schemes will compute the same answer (i.e., the diagram of Figure 2 commutes[4]).

## 1.3 Abstract Problem of Diagnosis

Diagnosis is the 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 an artifact.

The proposals to formalise the notion of diagnostic reasoning have generally considered two extremes of the diagnosis problem:

1. There is knowledge about how components are structured and work normally. There is no knowledge as to how malfunctions occur and manifest themselves. Diagnosis consists of isolating deviations from normal behaviour. This has normally been the preserve of consistency-based[5] approaches [7, 6].

2. There is knowledge about faults (diseases) and their symptoms, and we want to account for abnormal observations. This has traditionally been the preserve of abductive approaches [21, 22, 19, 4, 15].

[3] In the consistency-based diagnosis literature the result and the process are both called "diagnosis". Here, by "diagnosis" we mean the solution to the abstract diagnosis problem, not a representation of the problem or a representation of the solution.

[4] Here I only mean the solid lines. Of course, whether they compute what we really want to compute (i.e., whether the diagram with the light arrow also commutes) is a matter of argument, not of mathematics (see e.g., [20]).

[5] This term and "abduction" are used as technical terms defined in section 2.

In this paper, we consider fault based systems (as do [14, 2, 9]). In [16] we consider how the two logic-based models of diagnosis can use each sort of knowledge, and the continuum of cases between the two extremes.

Any diagnosis system requires knowledge about the domain of diagnosis and observations of the actual artifact we are diagnosing.

# 2 Two Models of Diagnosis

In this paper we cast two models of diagnosis into the Theorist framework of hypothetical reasoning [19, 15]. This formalism is well suited to the task as both paradigms can be naturally represented in the simple formal framework.

Theorist [15] is defined as follows. A knowledge base $KB$ is a pair $\langle F, H \rangle$, such that $F$ is a set of closed formulae[6] (called the *facts*) and $H$ is a set of open formulae (called the *possible hypotheses*). A **scenario** of $\langle F, H \rangle$ is a set $D \cup F$ where $D$ is a set of ground instances of elements of $H$ such that $D \cup F$ is consistent. An **explanation** of formula $g$ from $\langle F, H \rangle$ is a scenario of $\langle F, H \rangle$ that logically implies $g$. An **extension** of $\langle F, H \rangle$ is the set of logical consequences of a maximal (with respect to set inclusion) scenario of $\langle F, H \rangle$. Where the $KB$ is understood from context we omit the phrases "of $\langle F, H \rangle$", etc.

**Definition 2.1 (Consistency-Based Diagnosis)** A **consistency-based diagnosis** is minimal set of abnormalities such that the observations are consistent with all other components acting normally [23].

In terms of the Theorist framework,

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

$H$ is the set of normality assumptions.

**A consistency-based diagnosis** corresponds to an extension (in particular, it is the set of abnormalities in an extension) [23, theorem 6.1].

[6] We assume the underlying logic is the first order predicate calculus. We follow the Prolog convention of variables being in upper case. A set of formulae represents the conjunction of the formulae.

**Definition 2.2 (Abductive Diagnosis)** An **abductive diagnosis** is a minimal set of assumptions which, with a set of background knowledge implies the observations [19, 15], and is consistent with the observations.

In terms of the Theorist framework,

$F$ is the domain model.

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

**An abductive diagnosis** is a minimal explanation of the observations.

The main difference is that, in abduction the diagnoses entail the observations, whereas in consistency based models the observations entail the (disjunct of the) diagnoses. As one would expect the sort of knowledge that has to be specified for each is different.

# 3 KR for Each Diagnostic Formalism

## 3.1 Causes and Symptoms

As part of the terminology for talking about domains, 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. As far as the KR framework is concerned, we want a domain that can be described in terms of causes and symptoms.

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 that 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.

A "base cause" is a cause which don't need any further explanation (it is up to the user to determine what these are). An "observed symptom"

(or just "observation") is a symptom that we actually have observed. In particular base causes do not have further causes.

I also assume that there are no causal cycles. That is there is no causal chain from one proposition that goes back to itself. For example, it is never the case that $a$ causes $b$ and $b$ causes $a$. This is reasonable if we consider that the propositions represent particular events rather than event types. For example, consider a causal chain that "being stressed" causes one to "not work efficiently", which in turn causes one to "be stressed". We represent being stressed at different stages as different propositions that refer to different times. Being stressed in the past causes us to not work well at the moment which causes us to be stressed in the future. In terms of Lin's [10] causal dichotomy, we are talking about token causation, rather than type causation.

I mean something very different to Konolige's causal theories [9]. I do not mean a representation of causation but I mean causation itself. Whether or not causation can be represented in the way presented here (or even if causation is a property of the world) is an open question; it is not something that can be considered mathematically, but needs to be studies empirically by trying to represent (what purports to be) causation.

## 3.2 Fault Models

Consistency-based diagnosis is defined in terms of normality assumptions rather than in terms of fault (cause/symptom) models. Abductive diagnosis is conceptualised in terms of fault models. Before we can offer a detailed comparison, we have to consider how we could incorporate fault models into consistency-based diagnosis.[7]

To add fault models to consistency-based diagnosis, we need to address 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

[7]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 [23] is that an abnormality is whatever we are minimising.

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 [23] to model the generalised set covering model of [22]. In this paper I assume that this is the approach taken.

The diagnoses become the faults that can be proven from the assumption that other faults are absent [23, Proposition 3.3].

## 3.3 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 out what is causing the problems being manifested.

We first want to consider the question *what sort of knowledge is required?* At the top level of abstraction, to determine what sort of knowledge is required, we examine the definitions of the diagnostic paradigms to see what has to be proven.

1. In consistency-based diagnosis, we have to prove a fault[8] (maybe based on other assumptions) from an observation. Thus the sort of knowledge we need is of the form $\cdots \supset fault$.

2. In abductive diagnosis, the sort of knowledge we need is that from some explanation we can prove the observations. Thus the sort of knowledge we need is of the form $fault \supset symptoms$.

If $c_1, \cdots, c_n$ are the possible causes of symptom $s$, then for each of the paradigms we need to provide the following knowledge.

1. For consistency-based diagnosis we have as a fact $s \supset c_1 \vee \cdots \vee c_n$. If the artifact exhibits symptom $s$ then one of the causes of $s$ must be present. If $c_i$ always produces symptom $s$, then $s$ being false should rule out $c_i$; we should thus add the fact $c_i \supset s$ to the facts.

[8] Or equivalently, what follows from the negation of a fault. Note that the negation of a fault is the normality condition.

2. In abductive diagnosis, we have to be able to prove the symptoms from the causes. Thus the sort of knowledge is of the form $c_i \supset s$. If $s$ is always present when $c_i$ is present then $c_i \supset s$ should be a fact (the absence of $s$ can rule out $c_i$), otherwise $c_i \supset s$ should be a possible hypothesis (it can be used in an explanation, but not to rule out $c_i$).

**Example 3.1** Consider representing the following about how aching elbows and aching hands could be caused:

> *tennis-elbow* always causes *aching-elbow*.
> *dishpan-hands* sometimes causes *aching-hands*.
> *arthritis* sometimes causes *aching-elbow* and always causes *aching-hands*.

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

1. For consistency-based diagnosis, we can represent the above situation as

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

2. For abductive diagnosis, we can represent the above situation as

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

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

1. For the consistency-based diagnosis, 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. For the abductive diagnosis, there are two minimal explanations of *aching-elbow*:

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

Consider observing $\textit{aching-elbow} \wedge \textit{aching-hands}$.

1. For the consistency-based diagnosis, there are two extensions, one containing

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

and one containing

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

2. For the abductive diagnosis there are two minimal explanations of $\textit{aching-hands} \wedge \textit{aching-elbow}$:

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

This example can be very instructive on the differences between the diagnostic systems. The extensions of consistency-based diagnosis and the explanations of abductive diagnosis seem to be very similar (in Section 3.5 this equivalence is spelled out in greater detail).

## 3.4 Ruling out Causes

What sort of knowledge do we need to rule out consideration of particular causes? For example the knowledge that allows us to rule 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 (facts or defaults) of the form

$$evidence \supset \neg cause$$

These are "causal rules" because they give the implication of the symptoms from the causes. This is the sort of knowledge that abductive diagnosis needed in the first place, but is the opposite sort of implication than was claimed before to be needed in consistency-based diagnosis. Thus it seems as though in a system for consistency-based diagnosis one needs both causal rules and evidential rules.

Thus if $c_1, \cdots, c_n$ are the possible causes of $s$, then abductive diagnosis needs knowledge of the form

$$c_1 \supset s, \cdots, c_n \supset s$$

(those implications that are always true should be in $F$ and those causes that are not strict should be in $H$). Consistency-based diagnosis needs the strict implications as well as knowledge of the form

$$s \supset c_1 \vee \cdots \vee c_n$$

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 diagnostic paradigm can deviate from the standard case.

## 3.5 Standard Propositional case

The standard case we will consider places restrictions on the diagnostic problems we can represent:

1. The domain can be thought of in terms of causes and effects.

2. The domain can be described propositionally.

3. There is an acyclic causal structure. That is, if we write $c_i < s$ to mean atom $c_i$ is one of the causes for atom $s$, then the transitive closure of the binary relation $<$ is irreflexive.

From an understanding of this simple case, we can then learn about more complex cases. The first two assumptions are given up in [16]; the last assumption is given up in [9].

The base causes are those causes that themselves have no other causes. Unlike Konolige [9] we do not allow other causes of these base causes. If $c_i$ is some proposition that we would like to include as part of a diagnosis, but has some other cause, we cannot make $c_i$ into a base cause. One reason that we may want to make $c_i$ a base cause is if $c_i$ sometimes has other (known) causes, and sometimes $c_i$ may have no apparent (or represented) causes, and may just happen to be true. Instead of making $c_i$ a base cause, we create a new atom $c_i\_happens\_to\_be\_true$ and make it a base cause, and a cause for $c_i$. With this construction I argue that we would never want to make something imply what would otherwise be a base cause.

Suppose that for possible symptom (that is not a base cause) $s$, we have causes $c_1, \cdots, c_n$ (each of these can be a conjunction of base causes or even other non-base causes, which themselves have to be explained). If these causes are not covering we invent a new base cause *s_occurred_for_another_reason*, and add it to the set of possible causes. These new causes are now covering. We can thus assume, without loss of generality, that our set of causes is covering.

We also allow for integrity constraints of a quite general form. $C$ is a set of arbitrary propositional formulae such that if for some symptom $s$, we can derive $C \models w \supset s$, where $\not\models w \supset s$ and $c_i$ are the causes of $s$ then $C \models w \supset \bigvee_i c_i$. That is, if something non-trivially implies $s$ then it must imply some of the causes. This is a restriction on what we allow as the causes rather than what we allow as constraints.

We also assume that $C$ contains all implications of the form $c_i \supset s$ where $c_i$ always causes $s$ (and so the absence of symptom $s$ can be used to rule out $c_i$).

Let $B$ be the set of base causes, i.e., the causes that have no other causes. By the constraints on $C$, this means that there is no non-trivial formula $w$ such that $C \models w \supset b$ for $b \in B$.

We are now ready to define the corresponding knowledge bases.

Let the abductive knowledge base $KB_A$ be defined as

$$KB_A = \langle C, B \cup \{c_i \supset s : c_i \text{ causes } s, \text{ and } c_i \supset s \text{ is not in } C\}\rangle$$

Let the consistency-based knowledge base $KB_{CB}$ be defined as

$$KB_{CB} = \left\langle C \cup \{s \supset \bigvee_i c_i : \{c_i\} \text{ are the causes of } s\}, \{\neg b : b \in B\}\right\rangle$$

Thus if $c_1, \cdots, c_n$ are the possible causes of symptom $s$, then consistency-based diagnosis would represent this as $s \supset c_1 \vee \cdots \vee c_n$, and for each $c_i$ for which $s$ is a necessary symptom, we have $c_i \supset s$ as a fact. Abductive diagnosis would represent this as $c_i \supset s$ being a fact if $s$ is a necessary symptom of $c_i$, and $c_i \supset s$ as a possible hypothesis otherwise. Any other relationship between the two (e.g., a cause implying a disjunct of symptoms) would be added as facts to each of these.

**Theorem 3.2** Given a set of symptoms, the base causes in the diagnoses using abductive diagnosis from $KB_A$ are identical to the diagnoses using consistency-based diagnosis from $KB_{CB}$.

> **Proof:** We first prove this theorem for conjunctive queries (i.e., queries that are conjunctions of literals). We also assume that the knowledge base (the facts and each hypothesis) is in clausal form. As we can translate any formula into clausal form this places no restriction on the theory.
>
> The causal structure is acyclic and so forms a partial order. Define a total order consistent with this partial order, by assigning a natural number (called the *index*) to each atom such that the base causes have index zero and if atom $a$ is a cause of atom $b$, then the index of $a$ is less than the index of $b$. This can always be done as the causal structure forms a partial order with the base causes as the minimal elements.
>
> The theorem is proven by induction on the pair $\langle i, n\rangle$ where $i$ is an index and $n$ is the number of atoms in the observation that has index $i$, such that no atoms in the observation have an index greater than $i$. Each query can be associated with a pair.
>
> The base case for the induction is where either

1. $n = 0$, in which case the empty diagnosis is a diagnosis for each system, or
2. the maximum index of the observation is zero. In this latter case the observation is a conjunction of base causes. Suppose it is $b = b_1 \wedge \cdots \wedge b_n$. If $C \cup \{b\}$ is inconsistent there are no diagnoses in either system. Otherwise, in both systems the diagnosis is $b$.

For the inductive case, suppose that $s_1 \wedge \cdots \wedge s_n$ are our symptoms to be explained, with $s_1$ being an atom of maximal index. If $s_1$ is a base cause the induction can stop, as above. If it is not a base cause, there will be a (possibly empty) set of rules $c_i \supset s_1$ in $KB_A$ (facts or hypotheses). Consider the explanations of the "observation" $c_i \wedge s_2 \wedge \cdots \wedge s_n$ for each $i$. This is a query that is less in our inductive ordering, thus by the inductive assumption, the diagnoses from $KB_A$ and $KB_{CB}$ are identical. Suppose, that for each $i$, these are $D^i_1, \cdots, D^i_{k^i}$.

To make $D^i_j$ into an abductive diagnosis for $s_1, \cdots, s_n$ from $KB_A$, we have to

1. add $c_i \supset s_1$ to $D^i_j$ (if $c_i \supset s_1$ not in $C$),
2. check for consistency, and
3. check for minimality.

For the consistency based diagnosis we have

$$\begin{aligned} KB_{CB} &\models s_1 \supset \bigvee_i c_i \\ \therefore KB_{CB} &\models s_1 \wedge \cdots \wedge s_n \supset \bigvee_i c_i \wedge s_2 \wedge \cdots \wedge s_n \\ \therefore KB_{CB} &\models s_1 \wedge \cdots \wedge s_n \supset \bigvee_i \bigvee_j D^i_j \end{aligned}$$

The diagnoses of $s_1 \wedge \cdots \wedge s_n$ from $KB_{CB}$ consist of the subset of these that are consistent (as each diagnosis must prove all of the goals), and minimal. The important thing to notice is that exactly the same facts (i.e., those in $C$) are used to prune the

consistency-based diagnoses and the abductive diagnoses. The set of the $D^i_j$ that forms the set of preliminary diagnoses are pruned in exactly the same way for the abductive and consistency-based diagnoses, to form the same set of diagnoses.

□

It is important to note how the standard case works when there is no possible causes of a symptom. In the analysis above, for abductive diagnosis, this means that we cannot explain the symptom; for the representation for consistency-based diagnosis we have stated that the symptom could not occur (it implies the empty disjunction, which is false).

**Example 3.3** This example illustrates how the $D^i_j$ in the above proof need not be explanations of the observation. Suppose $c_1$ is a possible cause for $s$ and $c_2$ and $c_3$ are each possible base causes for $c_1$, and we have the constraint $c_2 \supset \neg s$. For this example

$$KB_A = \langle \{c_2 \supset \neg s\}, \{c_2, c_3, c_2 \supset c_1, c_3 \supset c_1, c_1 \supset s\} \rangle$$

$$KB_{CB} = \langle \{c_2 \supset \neg s, s \supset c_1, c_1 \supset c_2 \vee c_3\}, \{\neg c_2, \neg c_3\} \rangle$$

There are two diagnoses of $c_1$, namely $\{c_2\}$ and $\{c_3\}$. There is however only one diagnosis of $s$, namely $\{c_3\}$.

**Example 3.4** *Differences still arises if the knowledge is not of the form of our standard case. For example suppose the knowledge base contains $c_1 \vee c_2$, where $c_1$ and $c_2$ are base causes*[9] *and there are no observations. In abductive diagnosis, if there are no observations, then there is always the empty diagnosis if the knowledge base is consistent. For consistency-based diagnosis, 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 $c_1 \vee c_2$ as the knowledge base, there are two diagnoses ($\{c_1\}$ and $\{c_2\}$), even with no observations. Why and how one may want to exploit such distinctions is still an open question.*

[9] This violates our notion that nothing should imply a base cause. Here $\neg c_1 \supset c_2$.

## 3.6 Relationship to Clark's completion

If all causes are necessary causes, the sort of knowledge we need for abductive diagnosis is of the form

$$(c_1 \supset s) \wedge \cdots \wedge (c_n \supset s)$$

The sort of knowledge that we need for consistency-based diagnosis is of the form $s \supset c_1 \vee \cdots \vee c_n$ in order to conclude a cause, together with $c_i \supset s$ for each $i$ in order to rule out possible causes. Thus, it is of the form

$$s \equiv c_1 \vee \cdots \vee c_n$$

Notice that the second looks just like the completion (in terms of Clark [1]) of the first. In fact, it is closely related, but there are three 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. In general, the completion is with respect to our facts and hypotheses. We add the completion formula $a \supset c_1 \vee \cdots \vee c_n$, ignoring the distinction of whether each $c_i$ always causes $a$ or whether $c_i$ sometimes causes $a$. Thus we have to consider the implications in both the facts and the hypotheses. The causal implications in the hypotheses do not remain in the completion. We typically do not end up with a biconditional.

3. We are not only working with what [11] 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.

## 3.7 Pearl's example

**Example 3.5 (Pearl)** Pearl [13, p. 371] gives the following example 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 do not arise in consistency-based diagnosis and abductive diagnosis.

The situation 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.*

For consistency-based diagnosis, we would represent this situation as:

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

For abductive diagnosis, we would represent the same situation as

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

Suppose that we observe that it rained last night.
For the consistency-bead diagnosis, there is one extension containing

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

For the abductive diagnosis, there is one explanation of *rained-last-night*, namely

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

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

In Pearl's rule-based system [13], he can explain everything, including that the sprinkler was on last night. Pearl attributes this problem to not distinguishing between evidential and causal rules. I would claim that it is a flaw in the idea of rule-based diagnosis used by Pearl.

Suppose we had instead observed that the grass is cold and shiny.

For the consistency-bead diagnosis, there are two extensions,

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

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

For the abductive diagnosis, there are two explanations

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

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

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

The following example shows that can represent more than definite clauses:

**Example 3.6** Suppose we have three causes $c_1$, $c_2$ and $c_3$ and symptoms $s_1$ and $s_2$ such that:
$c_1$ always produces symptom $s_1$.
$c_2$ always produces either symptom $s_1$ or $s_2$.
$c_3$ sometimes produces symptom $s_2$.
$c_1$ and $s_2$ cannot co-occur.

For the abductive framework this is represented as:

$$\begin{aligned} F &= \{ \; c_1 \supset s_1, \\ &\qquad c_1 \supset \neg s_2, \\ &\qquad c_2 \supset s_1 \vee s_2 \} \\ H &= \{ \; c_2 \supset s_1, \\ &\qquad c_2 \supset s_2, \\ &\qquad c_3 \supset s_2, \\ &\qquad c_1, \\ &\qquad c_2, \\ &\qquad c_3 \} \end{aligned}$$

For the consistency-based diagnosis this is written as

$$\begin{aligned} F &= \{ \; c_1 \supset s_1, \\ &\qquad c_1 \supset \neg s_2, \end{aligned}$$

$$
\begin{aligned}
& c_2 \supset s_1 \vee s_2, \\
& s_1 \supset c_1 \vee c_2, \\
& s_2 \supset c_2 \vee c_3\} \\
H \quad = \{ \quad & \neg c_1, \neg c_2, \neg c_3\}
\end{aligned}
$$

These two theories have the same diagnoses.

Note that we do not use negation as failure for explanation. If we had observed $\neg s_1$, then either we would have to have causes for $\neg s_1$, or it would have to be a base cause in order for there to be diagnoses for an observation including $\neg s_1$. This can be done in the framework presented here, and is, for example, systematically done in [18].

# Conclusion

In this paper we have presented an abstract knowledge representation problem applied to diagnosis. To understand the technical results of this paper, it is important to understand them in the context of the knowledge representation problem presented in Section 1.1. This is important in that there may be many different ways to represent a problem. Some restrictions placed on the results in this paper are restrictions in the knowledge base and not in what can be represented (e.g., the fact that base causes have no other causes), whereas other restrictions are restrictions in what can be represented (e.g., acyclicity of the causal structure).

This paper should be seen as a starting point for understanding the knowledge representation issues in diagnosis. For example, in understanding more complex diagnostic domains [16, 17], and representing uncertainty in diagnosis [18].

# Acknowledgements

This research was supported under NSERC grant OGPOO44121, and under Project B5 of the Institute for Robotics and Intelligent Systems.

# References


[1] K. L. Clark. Negation as failure. In H. Gallaire and J. Minker, editors, *Logic and Databases*, pages 293–322. Plenum Press, New York, 1978.

[2] L. Console, D. Theseider Dupre, and P. Torasso. Abductive reasoning through direct deduction from completed domai[225zn models. In W. R. Zbigniew, editor, *Methodologies for Intelligent Systems 4*, pages 175–182. Elsiever Science Publishing Co., 1989.

[3] L. Console, D. Theseider Dupre, and P. Torasso. On the relationship between abduction and deduction. *Journal of Logic and Computation*, 1(5):661–690, 1991.

[4] P. T. Cox and T. Pietrzykowski. General diagnosis by abductive inference. Technical Report CS8701, Computer Science, Technical University of Nove Scotia, Halifax, April 1987.

[5] E. Davis. *Representations of Commonsense Knowledge*. Morgan Kaufmann, San Mateo, Cal., 1990.

[6] J. de Kleer and B. C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32(1):97–130, April 1987.

[7] M. R. Genesereth. The use of design descriptions in automated diagnosis. *Artificial Intelligence*, 24(1-3):411–436, December 1984.

[8] M. L. Ginsberg, editor. *Readings in Nonmonotonic Reasoning*. Morgan Kaufmann, Los Altos, Cal., 1987.

[9] K. Konolige. Abduction versus closure in causal theories. *Artificial Intelligence*, 53(2-3):255–272, February 1992.

[10] D. Lin. A probabilistic network of predicates. In D. Dubois, M. P. Wellman, B. D'Ambrosio and P. Smets, editor, *Proc. Eighth Conf. on Uncertainty in Artificial Intelligence*, pages 174–181, Stanford University, July 1992.

[11] J. W. Lloyd. *Foundations of Logic Programming*. Symbolic Computation Series. Springer-Verlag, Berlin, second edition, 1987.

[12] J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In M. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969.

[13] J. Pearl. Embracing causation in default reasoning. *Artificial Intelligence*, 35(2):259–271, 1988.

[14] D. Poole. Representing knowledge for logic-based diagnosis. In *International Conference on Fifth Generation Computing Systems*, pages 1282–1290, Tokyo, Japan, November 1988.

[15] D. Poole. Explanation and prediction: an architecture for default and abductive reasoning. *Computational Intelligence*, 5(2):97–110, 1989.

[16] D. Poole. Normality and faults in logic-based diagnosis. In *Proc. 11th International Joint Conf. on Artificial Intelligence*, pages 1304–1310, Detroit, August 1989.

[17] D. Poole. A methodology for using a default and abductive reasoning system. *International Journal of Intelligent Systems*, 5(5):521–548, December 1990.

[18] D. Poole. Probabilistic Horn abduction and Bayesian networks. Technical Report 92-20, Department of Computer Science, University of British Columbia, August 1992. To appear, *Artificial Intelligence* 1993.

[19] D. Poole, R. Goebel, and R. Aleliunas. Theorist: A logical reasoning system for defaults and diagnosis. In N. Cercone and G. McCalla, editors, *The Knowledge Frontier: Essays in the Representation of Knowledge*, pages 331–352. Springer-Verlag, New York, NY, 1987.

[20] D. Poole and G. Provan. What is the most likely diagnosis? In P. P. Bonissone, M. Henrion, L. N. Kanal and J. F. Lemmer, editor, *Uncertainty in Artificial Intelligence 6*, pages 89–105. Elsevier Science Publishers B. V., 1991.

[21] H. E. Pople, Jr. On the mechanization of abductive logic. In *Proc. 3rd International Joint Conf. on Artificial Intelligence*, pages 147–152, Stanford, August 1973.

[22] J. Reggia, D. Nau, and P. Wang. A formal model of diagnostic inference. *Information Sciences*, pages 227–285, 1985.

[23] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95, April 1987.