Symbolic Diagnosis and its Formalisation[*]

Peter Lucas
Department of Computer Science, Utrecht University
P.O. Box 80.089
3508 TB Utrecht, The Netherlands
(e-mail: lucas@cs.ruu.nl)

Abstract

Diagnosis was among the first subjects investigated when digital computers became available. It still remains an important research area, in which several new developments have taken place in the last decade. One of these new developments is the use of detailed domain models in knowledge-based systems for the purpose of diagnosis, often referred to as model-based diagnosis. Typically, such models embody knowledge of the normal or abnormal structure and behaviour of the modelled objects in a domain. Models of the structure and workings of technical devices, and causal models of disease processes in medicine are two examples. In this article, the most important notions of diagnosis and their formalisation are reviewed and brought in perspective. In addition, attention is focused on a number of general frameworks of diagnosis, which offer sufficient flexibility for expressing several types of diagnosis.

Keywords: diagnostic expert systems, model-based diagnosis, theory of diagnosis.

Contents

[*] This paper has been published in: *The Knowledge Engineering Review*, **12**(2), 109–146, 1997.

1 Introduction

Diagnosis is commonly viewed as the interpretation of case-specific findings in the context of knowledge from a problem domain to obtain an indication of the presence and absence of defects or faults, and also of the nature of the problem. Computer-aided diagnosis was among the first applications investigated when digital computers became available more than four decades ago. It still remains an important research area, in which several new developments have taken place in the last decade. Diagnosis is the subject of this review article.

It is customary to distinguish between diagnostic systems based on symbolic, or qualitative, reasoning technology, and those based on probability theory and statistics, although some systems offer a mixture of the two approaches. In this paper, we focus on systems based on symbolic reasoning technology.

It is, perhaps, not surprising that medicine was one of the first areas in which diagnostic expert systems were developed. Classical diagnostic medical expert systems are: INTERNIST-1 and its commercially available successor QMR, expert systems in the broad domain of internal medicine [Miller et al., 1982; Bankowitz et al., 1989], CASNET, an expert system for the diagnosis and treatment of glaucoma [Kulikowski & Weis, 1982; Weiss et al., 1978], ABEL, an expert system for the management of electrolyte and acid-base derangements [Patil, 1981; Patil et al., 1982], and the well-known MYCIN system, an expert system for the diagnosis and treatment of septicaemia and meningitis [Buchanan & Shortliffe, 1984; Shortliffe, 1976]. Many diagnostic systems have also been developed in technical fields, solving a wide variety of diagnostic problems. Early examples of such systems are SPERIL, a system that assisted in assessing damage to buildings after an earthquake [Ishizuka et al., 1981], and CRIB, a system that assisted in diagnosing computer hardware faults [Johnson & Keravnou, 1988].

Although all systems mentioned above can be viewed as knowledge-based systems for diagnosis, they are actually based on different, sometimes related, principles, as is apparent from the descriptions available in the literature (e.g. [Buchanan & Shortliffe, 1984], [Clancey & Shortliffe, 1984], [Johnson & Keravnou, 1988], [Szolovits, 1982] contain extensive descriptions of several systems). Until recently, however, no theoretical framework was available to formally describe and compare the various underlying principles. At a conceptual level, it was evident that the knowledge bases of some of the systems captured models of structure and behaviour in a domain. Such systems have been called *model-based* or '*first principles*' systems [Weiss et al., 1978; Davis, 1983]. The knowledge bases of other systems, however, did not embody an explicit model of structure and behaviour, but rather consisted of encoded human expertise in solving particular problems in the underlying domain. Currently, the term *empirical associations* is often employed to denote such knowledge. The classical example of such a system is MYCIN [Shortliffe, 1976].

The model-based approach to diagnosis has been successfully applied to fault finding in electronic circuits. Early work in this field is described in [Brown et al., 1982], [Davis, 1984], [Genesereth, 1984] and [De Kleer, 1976]. The study of simple electronic circuits has yielded much insight into the nature of the diagnostic process. More importantly, one of the first formal theories of diagnosis emerged from this research: the theory of consistency-based diagnosis as proposed by R. Reiter [Reiter, 1987]. *Consistency-based diagnosis* offers a logic-based framework to formally describe diagnosis of abnormal behaviour in a device or system, using a model of normal structure and functional behaviour. Basically, consistency-based diagnosis amounts to finding faulty device components that account for a discrepancy between predicted normal device behaviour and observed (abnormal) behaviour. The predicted behaviour is inferred from a formal model of normal structure and behaviour of the device.

The systems CASNET and ABEL, mentioned above, are other early examples of model-based systems, but the principles of these systems differ from those used for technical devices. In particular, consistency-based diagnosis is not very suitable to describe diagnostic problem solving in these systems. Both CASNET and ABEL contain a representation of disease progress in terms of cause-effect (causal) relationships. In a sense these cause-effect relationships capture the 'behaviour' of disease processes. Causal knowledge is also incorporated in recent systems like the Oxford System of Medicine (OSM), [Fox et al., 1990b], a large medical expert system for general practitioners, and the DILEMMA toolset, [Huang et al., 1993], and generally seems to gain in importance in knowledge engineering.

Where consistency-based diagnosis traditionally employs a model of *normal* behaviour, *abduction* has been the principal model-based technique for describing and analysing diagnosis using a model of *abnormal* behaviour in terms of cause-effect relationships [Console et al., 1989; Console & Torasso, 1990a; Josephson & Josephson, 1994; Reggia et al., 1983; Peng & Reggia, 1990; Poole, 1988; Wu, 1991]. Early work on abduction has been done by H.E. Pople (cf. [Pople, 1973; Pople, 1977]) and D. Poole (cf. [Poole et al., 1987]). Some of the early diagnostic systems that incorporated causal knowledge, such as ABEL and CASNET, are nowadays viewed as being, at least partially, abductive in nature. In *abductive diagnosis*, diagnostic problem solving consists of establishing a diagnosis using cause-effect relationships with a set of observed findings (effects) as the starting point. In abduction, a system reasons from effects to causes, instead of from causes to effects. Because the reasoning from causes to effects can be accomplished using logical deduction, in a sense abductive reasoning is carried out in a direction reverse to that of deduction.

Logical deduction, however, also has its place in the picture, because it has been used to formalise reasoning with the logical analogues of empirical associations [Lucas, 1993]. In the context of diagnosis, reasoning with empirical associations is often referred to as *heuristic classification* [Clancey, 1985].

Although much work has now been done to formalise diagnosis, it has been difficult to capture the concept of diagnosis in a precise, formal and also general way, leaving room for various types of diagnosis. Both consistency-based diagnosis and abductive diagnosis have been looked upon as core concepts for formal frameworks of diagnosis, but, as we shall see, other formalisations are also possible. A formal framework of diagnosis offers means to formally describe and analyse various notions of diagnosis. The frameworks described in the literature are either logic-based (cf. for example [Console et al., 1991], [Konolige, 1994], [Poole, 1990b; Poole, 1994], and [Ten Teije & Van Harmelen, 1994]) or based on set theory (cf. for example [Josephson & Josephson, 1994] and [Lucas, 1996a]).

The formalisation of diagnosis is the subject reviewed in this article. The structure of this article is as follows. First, the nature of the diagnostic process is sketched. Next, the various core approaches to diagnosis described in the literature are reviewed. Finally, the various approaches to diagnosis are compared to each other, and a number of frameworks that offer means for the general description of diagnosis are discussed.

2 Diagnostic problem solving

Discovering what is wrong in a particular situation is one of the central activities in real life; this process is usually called *diagnosis* or *diagnostic problem solving.* The process may be viewed as the selective gathering and interpretation of information as evidence for or against the presence or absence of one or more defects in a system. This informal definition reveals that the following aspects are of central importance to diagnostic problem solving. Firstly, the *gathering* of information, and secondly, the *interpretation* of the gathered information for determining what is wrong, for example with a patient or a device. In medicine, defects are disorders of a patient; in technical domains, defects are faults of a device. In medicine, the information-gathering process is usually carried out in a systematic, structured fashion, because there are an enormous number of diagnostic tests available to the clinician, that cannot all be carried out. Furthermore, some diagnostic tests cause discomfort to the patient, or carry even some risk of causing disease or death. By restricting the selection of diagnostic tests in early diagnosis to those that do no harm or cause little discomfort to the patient, as is common practice in medical diagnosis, diagnostic tests are performed only when necessary. In technical fields, it is sometimes impossible to gather certain information because of time constraints, costs involved, or physical impossibility. Although the information-gathering process is a characteristic feature of diagnosis, the interpretation of information as evidence for or against a diagnostic solution is a more fundamental aspect of diagnostic problem solving.

The information-gathering process together with related aspects, such as the process of generating, and accepting or rejecting diagnostic hypotheses are sometimes referred to as the *dynamic* aspects of diagnostic problem solving. They yield specific problem-solving behaviour. Establishing an actual diagnostic solution requires knowledge of what constitutes a diagnosis of a particular problem; the various aspects involved are sometimes referred to as the *static* aspects of diagnostic problem solving. This article focuses on these static aspects of diagnostic problem solving.

In general, diagnostic problem solving, like many other forms of problem solving, may be described using the scientific notion of the *empirical cycle*, which describes the framework underlying empirical research [Popper, 1959]. It states that empirical research encompasses: (1) formulating a *hypothesis*, (2) *testing* that hypothesis, and (3) *rejecting* the hypothesis when it fails to pass the tests, or *accepting* the hypothesis when it successfully passes the tests.[1] The process may start again with (1), in which case the formulation of a new hypothesis possibly involves *adjusting* a hypothesis previously rejected. In Figure 1, this view of diagnostic problem solving as an instance of the empirical cycle is depicted. Testing involves the application of procedures for the verification and falsification of a hypothesis using *observed findings* and domain knowledge. In general, a hypothesis may be a complex structure or mechanism. In diagnostic problem solving, however, a hypothesis is usually taken to be a collection of 'defects', where each defect is assumed to be either present or absent. This simplification may not always be justified, for example because the defects may be interrelated to each other in some particular way, which could be part of the hypothesis. For example, a hypothesis may be whether or not a process A is causally related to a process B. Nevertheless, this simplification is invariably made in diagnostic systems, and seems acceptable in the light of developed applications. A *diagnosis* may be conceived as an accepted hypothesis concerning a particular defect or collection of defects; the results of diagnostic tests correspond to the

[1] Popperians may read instead: 'not rejecting the hypothesis so long as it has not been falsified by a test'.

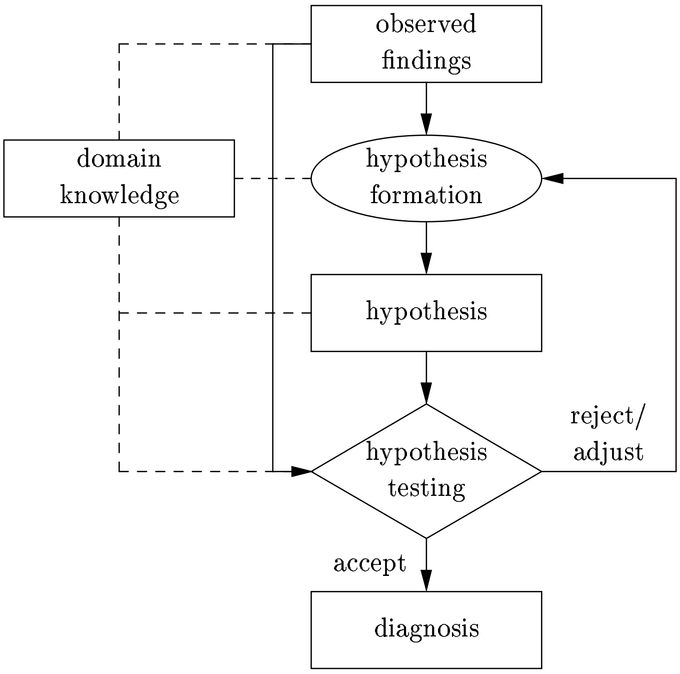

Figure 1: Diagnostic problem solving and the empirical cycle.

observed findings.

The literature on diagnosis more or less follows the terminology and structure of the empirical cycle. For example, [Davis & Hamscher, 1988] views diagnostic problem solving as three fundamental subproblems:

(1) Hypothesis generation (or hypothesis formation);

(2) Hypothesis testing;

(3) Hypothesis discrimination.

The subproblem of hypothesis discrimination concerns selecting from the hypotheses accepted on the basis of a measure of plausibility. This process may entail collecting additional observed findings.

The basic framework of diagnostic problem solving as the empirical cycle can be refined in several ways. For example, there may be an ordering on the set of hypotheses, such as an ordering from generic to specific, or an ordering by the value of a real-valued utility function associated with the hypotheses. A class of defects may be taken as a generic hypothesis, and a specific defect may be viewed as a specific hypothesis. Such orderings are especially useful in guiding the problem-solving process, information gathering included. For example, the process may be decomposed into several stages working from generic towards more specific hypotheses, or from hypotheses with high associated utility to those with low associated utility. It is well-known that guiding the problem-solving process, using information collected at earlier stages, may be quite effective in reducing the number of tests to be performed. It may also result in a step-wise reduction in the number of defects to be considered, due to the rejection of specific hypotheses motivated by the earlier rejection of more generic hypotheses. This approach to handling hypotheses and observable findings is an example of a so-called (diagnostic) *problem-solving strategy* [Newell & Simon, 1972]. Problem-solving strategies are beyond the scope of this article, because they belong to the dynamics of diagnosis.

3 Conceptual basis of diagnosis

Although the description of diagnostic problem solving given in Section 2 carries much of the flavour of the process of diagnosis, it is still an imprecise description and, in fact, several formal theories have been proposed to capture the concept of diagnosis more precisely. In doing so, however, researchers became aware that there are actually various *conceptual models* of diagnosis, determined by the kind of knowledge involved. As stated in Section 1, diagnosis concerns the interpretation of observed findings in the context of knowledge from a problem domain. A good starting point for describing diagnosis at a conceptual level are the various types of knowledge that play a role in diagnostic applications.

The knowledge embodied in a diagnostic system may be based on one or more of the following descriptions:

(1) A description of the *normal* structure and functional behaviour of a system.

(2) A description of *abnormal* functional behaviour of a system; abnormal structure is usually not taken into account.

(3) An enumeration of defects and collections of observable findings for every possible defect concerned, without the availability of explicit knowledge concerning the (abnormal) functional behaviour of the system.

(4) An enumeration of findings for the normal situation.

These types of knowledge may coexist in real-life diagnostic systems, but it is customary to emphasise their distinction in conceptual and formal theories of diagnosis. Similar classifications of types of knowledge appear in the literature on diagnosis, although often no clear distinction is made between the conceptual, formal and implementation aspects of diagnostic systems. For example, [Davis & Hamscher, 1988] and [Poole, 1988] distinguish diagnostic rule-based systems, by which they mean diagnostic systems based on knowledge of the third type mentioned above, from diagnostic systems incorporating knowledge of structure and behaviour, i.e. knowledge of the first and second type mentioned above. However, rule-based systems with a sufficiently expressive production-rule formalism can be used to implement any diagnostic system, including those based on knowledge of structure and behaviour.

An observed finding that has been gathered in diagnosing a problem is often said to be either a ‘normal finding’, i.e. a finding that matches the normal situation, or an ‘abnormal finding’, i.e. a finding that does not match the normal situation. Based on the four types of knowledge mentioned above, and the two sorts of findings, three different conceptual models of diagnosis are usually distinguished; they will be called:

- *Deviation-from-Normal-Structure-and-Behaviour diagnosis*, abbreviated to *DNSB diagnosis*,
- *Matching-Abnormal-Behaviour diagnosis*, abbreviated to *MAB diagnosis*, and
- *Abnormality-Classification diagnosis*, abbreviated to *AC diagnosis.*

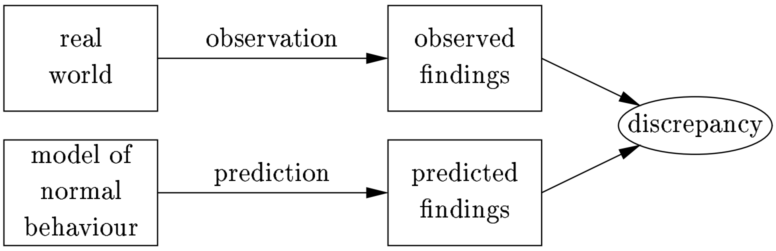

Figure 2: Deviation-from-normal-structure-and-behaviour (DNSB) diagnosis.

Below, we shall discuss the relationship between these three conceptual models of diagnosis and the four types of knowledge mentioned above. A formal theory of diagnosis has been proposed for each of these conceptual models of diagnosis. In the remainder of this section, each of the three conceptual models of diagnosis will be discussed, and the corresponding formal theory of diagnosis is mentioned. The formal theories of diagnosis are discussed in depth in Section 4.

DNSB diagnosis. For diagnosis based on knowledge concerning normal structure and behaviour, little or no explicit knowledge is available about the relationships between defects of the system, on the one hand, and findings to be observed when certain defects are present, on the other hand. Hence, DNSB diagnosis typically employs knowledge of the first and fourth types mentioned above. From a practical point of view, the primary motivation for investigating this approach to diagnosis is that in many domains little knowledge concerning abnormality is available, which is certainly true for new human-developed artifacts. For example, for a new device that has just been released from the factory, experience with respect to the faults that may occur when the device is in operation is lacking. Thus, the only conceivable way in which initially such faults can be handled is by looking at the normal structure and functional behaviour of the device. Yet, even if knowledge concerning abnormal behaviour is available, exhaustive description may be sometimes too cumbersome compared with a model of normal behaviour.

For the purpose of diagnosis, the actual behaviour of a physical device, called *observed behaviour*, is compared with the results of a model of normal structure and behaviour of the device, which may be taken as *predicted behaviour*. Both types of behaviour can be characterised by findings. If there is a *discrepancy* between the observed and the predicted behaviour, diagnostic problem solving amounts to isolating the components in the device that are not properly functioning, using a model of the normal structure and behaviour of the device [Brown et al., 1982; Davis, 1984; Davis & Hamscher, 1988; Genesereth, 1984; De Kleer, 1976]. In doing so, it is assumed that the model of normal structure and behaviour is sufficiently accurate and correct. Figure 2 depicts DNSB diagnosis in a schematic way. DNSB diagnosis is frequently erroneously called model-based diagnosis in the literature, as if it were the only instance of model-based diagnosis. It is also called consistency-based dagnosis, but in this article this term is reserved for the corresponding formal theory of diagnosis. DNSB diagnosis has been developed in the context of troubleshooting in electronic circuits [Davis & Hamscher, 1988]. A well-known program that supports DNSB diagnosis, and includes various strategies to do so efficiently, is the *General Diagnostic Engine* (GDE) [De Kleer & Williams, 1987; Forbus & De Kleer, 1993].

Above, we have reviewed the conceptual basis of diagnosis based on a model of normal

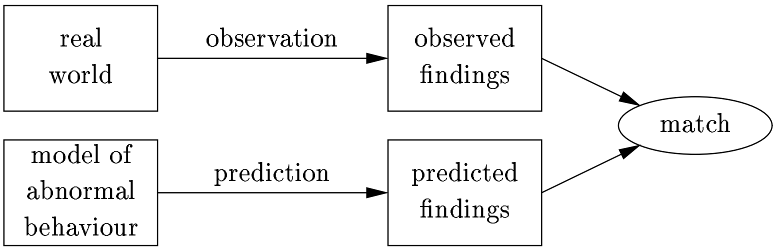

Figure 3: Matching-abnormal-behaviour (MAB) diagnosis.

structure and behaviour, which we have called DNSB diagnosis. The formal counterpart of DNSB diagnosis, called *consistency-based diagnosis*, originates from work by R. Reiter, [Reiter, 1987]; consistency-based diagnosis will be discussed in detail below. As far as known to the author, DNSB diagnosis-like approaches have been used in medical applications on a limited scale (cf. for example [Downing, 1993]); there is more work in which DNSB diagnosis has been applied to solve technical problems (cf. [Beschta et al., 1993; Dague, 1994; Hamscher, 1994; Ng, 1991; Sauthier & Faltings, 1992; Stefanini et al., 1993]).

MAB diagnosis. For diagnosis based on knowledge of abnormal behaviour, diagnostic problem solving amounts to simulating the abnormal behaviour using an explicit model of that behaviour. Hence, in MAB diagnosis the use of knowledge of abnormal behaviour (the second type mentioned above) is emphasised. By assuming the presence of certain defects, some observable abnormal findings can be predicted. It can be investigated which of these assumed defects account for the observed findings by *matching* the predicted abnormal findings with those observed. In Figure 3, MAB diagnosis is depicted schematically. In most applications of MAB diagnosis, the domain knowledge that is used for diagnosis consists of causal relationships. Two, strongly related, formal counterparts of MAB diagnosis have been proposed in the literature. The first formal theory, referred to as the *set-covering theory of diagnosis*, is based on set theory: causal knowledge is expressed as mathematical relations, used for diagnosis. This theory originates from work by J.A. Reggia and others [Reggia et al., 1983]. The second theory is based on logic. Early work in this area has been done by P.T. Cox and T. Pietrzykowski [Cox & Pietrzykowski, 1987], D. Poole [Poole et al., 1987], and by L. Console and P. Torasso [Console et al., 1989; Console & Torasso, 1990a]. Based on the type of reasoning employed to formalise MAB diagnosis, i.e. reasoning from effects to causes instead of from causes to effects, this theory of diagnosis is also referred to as *abductive diagnosis*. Theorist [Poole et al., 1987; Poole, 1990c] and CHECK [Console & Torasso, 1989] are two systems supporting MAB diagnosis.

AC diagnosis. Where DNSB and MAB diagnosis employ a model of normal or abnormal structure and behaviour for the purpose of diagnosis, the third conceptual model of diagnosis uses neither. The knowledge employed in this conceptual model of diagnosis consists of the enumeration of more or less typical evidence that can be observed, i.e. observable findings, when a particular defect or defect category is present (the third type of knowledge mentioned above). For example, sneezing is a finding that may be typically observed in a disorder like common cold. This form of knowledge has been referred to as *empirical associations* in Section 1 (the phrase '*compiled knowledge*' is also employed) [Buchanan & Shortliffe, 1984]. Diagnostic problem solving amounts to establishing which of the elements in a finite set of

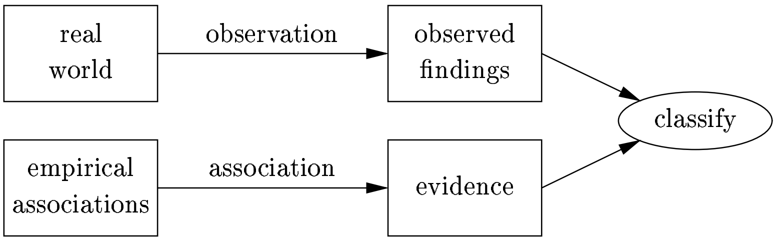

Figure 4: Abnormality-classification (AC) diagnosis.

defects have associated findings that account for as many of the findings observed as possible, as is shown in Figure 4. The enumeration of findings for the normal situation (knowledge of the fourth type mentioned above) is sometimes also used in AC diagnosis, together with knowledge of the third type; then, observed findings are classified in terms of present and absent defects. The main goal of AC diagnosis, however, remains the classification of observed findings in terms of abnormality. AC diagnosis is often referred to in the literature as *heuristic classification* [Clancey, 1985], although this term is broader, since it also includes a reasoning strategy. The MYCIN system, [Shortliffe, 1976], is the classical system in which this conceptual approach to diagnosis has been adopted. AC diagnosis can be characterised in terms of logical deduction in a straightforward way. We shall refer to this formalisation of AC diagnosis as *hypothetico-deductive diagnosis*.

A comparison of the three conceptual models of diagnosis is given in Table 1. Obviously, the various models of diagnosis discussed above can also be combined. To solve real-life diagnostic problems in a domain, it is likely that a mixture of conceptual models of diagnosis as distinguished above will be required. Since the resulting systems use various types of knowledge, e.g. both knowledge of structure and behaviour, and empirical associations, the result is known as diagnosis with *multiple models* [Struss, 1992]. Several programs have been developed that offer limited possibility to carry out diagnostic problem solving using multiple models; examples of such programs are GDE$^+$ [De Kleer, 1977; Struss & Dressler, 1989] and Sherlock [De Kleer & Williams, 1989]. These programs use DNSB diagnosis as their core approach.

Although in the literature it is emphasised that the conceptual models of diagnosis discussed embody different forms of diagnosis, they have much in common. For example, the type of knowledge used in DNSB diagnosis can be viewed as an implicit, or intensional, version of the type of knowledge used in AC diagnosis (if restricted to normality classification), which is an explicit or extensional type of knowledge; the associations between normal observable

	DNSB	**MAB**	**AC**
Type of knowledge	normal structure and behaviour	causal model of abnormality	empirical associations
Formalisation	consistency-based diagnosis	abductive and set-covering diagnosis	hypothetico-deductive diagnosis
Examples of systems	GDE	Theorist/CHECK	EMYCIN

Table 1: Comparison of typical conceptual models of diagnosis.

findings and the absence of defects are hidden in the specified normal behaviour in DNSB diagnosis. DNSB and MAB diagnostic problem solving are based on some kind of simulation of behaviour; such simulation of behaviour is absent in AC diagnosis.

4 Formal theories of diagnosis

There have been several attempts to formalise the various conceptual models of diagnosis discussed above; most, but not all, of these formalisations are based on logic. The most important formal theories of diagnosis will be reviewed below.

4.1 Consistency-based diagnosis

The formal theory of diagnosis originally proposed by R. Reiter, [Reiter, 1987], was motivated by the desire to provide a formal underpinning of diagnostic problem solving using knowledge of the normal structure and behaviour of technical devices, i.e. DNSB diagnosis. The theory of diagnosis may be viewed as the logical foundation of earlier work in DNSB diagnosis by J. de Kleer et al. [De Kleer, 1976; De Kleer & Williams, 1987], Brown and colleagues [Brown et al., 1982], R. Davis and H. Shrobe [Davis & Shrobe, 1983; Davis, 1984], and M.R. Genesereth [Genesereth, 1984]. The logical formalisation uses results from earlier work by R. Reiter, [Reiter, 1980], and J. McCarthy, [McCarthy, 1986], on nonmonotonic reasoning. We shall sometimes refer to this theory of diagnosis as Reiter's formal theory of diagnosis.

Reiter's theory of diagnosis was later extended by De Kleer et al. [De Kleer et al., 1992]; in this section, both formalisations will be introduced in a single, logical framework. Where appropriate, the differences between Reiter's original proposal, [Reiter, 1987], and the extensions proposed in [De Kleer et al., 1992] will be indicated. This formal theory of diagnosis is often referred to as the *consistency-based theory of diagnosis*, or *consistency-based diagnosis* for short.

The logical specification of knowledge concerning structure and behaviour in Reiter's theory is a triple $\mathcal{S} = (\mathrm{SD}, \mathrm{COMPS}, O)$, called a *system*, where

- SD denotes a finite set of formulae in first-order predicate logic, specifying normal structure and behaviour, called the *system description*;
- COMPS denotes a finite set of constants (nullary function symbols) in first-order logic, denoting the *components* of the system;
- O denotes a finite set of formulae in first-order predicate logic, denoting *observations*, i.e. observed findings.

It is, in principle, possible to specify normal as well as abnormal (faulty) behaviour within a system description SD, but originally SD was designed to comprise a logical specification of normal behaviour of the modelled system only, thus yielding the intended formalisation of DNSB diagnosis. The essential part of a formal model of normal structure and behaviour of a system consists of logical axioms of the form

$$\neg Abnormal(c) \to o_{norm} \tag{1}$$

where $c \in \mathrm{COMPS}$, and o_{norm} denotes a finding that may be observed if the component c is normal, i.e. is nondefective. The observable finding o_{norm} need not be unique. Axioms of the above form are provided for each component $c \in$ COMPS. These axioms will be referred to as *normality axioms.* It is assumed that the finding o_{norm} may be observed in reality when component c of the device, that has been modelled in logic, is operating normally. Such an observed finding is called a *normality observation.* The subscript *norm* is used to emphasise that a particular finding represents a normal result; in Section 4.2 and further, the subscript *ab* is used to indicate an abnormal finding. These subscripts are only used for clarity and have no additional meaning; they will often be omitted. The predicate symbol '*Abnormal*' is sometimes referred to as the *fault mode* (also behavioural mode) of the component [De Kleer & Williams, 1989]. The literal '$Abnormal(c)$' denotes the component c to be defective if satisfied. Other predicate names, such as 'OK', '*Correct*', are also employed in the literature, with similar intended meaning and use as the negation of an '*Abnormal*' literal.

Diagnostic problem solving is formalised as a method for finding the source of inconsistency in the logical description of the (normal) functioning of a system when supplied with observed findings, where some of the observed findings are the result of a system defect in reality. Hence, inconsistency formalises the notion of discrepancy in DNSB diagnosis as indicated in Figure 2. If it is assumed that the atom $Abnormal(c)$ is *false*, i.e. the component c is functioning normally, inconsistency will arise given the observed finding $\neg o_{norm}$ with logical implication (1). This result is interpreted in Reiter's theory as an indication that the defect may be localised in component c. This gives rise to the hypothesis that component c is defective, i.e. $Abnormal(c)$ is *true*, and the inconsistency is resolved if the assumption that $Abnormal(c)$ is *false* was its only source. In [Davis, 1984], this effect of relaxing logical constraints is referred to as *constraint suspension.*

Adopting the definition from [De Kleer et al., 1992], a diagnosis in the theory of consistency-based diagnosis can be defined as follows.

Definition 1 (*consistency-based diagnosis*)**.** *Let* $\mathcal{S} = (\mathrm{SD}, \mathrm{COMPS}, O)$ *be a system. Let*

$$H_P = \{Abnormal(c) \mid c \in \mathrm{COMPS}\}$$

be the set of all positive 'Abnormal' literals, and

$$H_N = \{\neg Abnormal(c) \mid c \in \mathrm{COMPS}\}$$

be the set of all negative 'Abnormal' literals. Furthermore, let $H \subseteq H_P \cup H_N$ *be a set, called a* hypothesis, *such that*

$$H = \{Abnormal(c) \mid c \in D\} \cup \{\neg Abnormal(c) \mid \mathrm{COMPS}\backslash D\}$$

for some $D \subseteq$ COMPS. *Then, the hypothesis* H *is a* (consistency-based) diagnosis *of* $\mathcal{S}$ *if the following condition, called the* consistency condition, *holds:*

$$\mathrm{SD} \cup H \cup O \nvDash \bot \tag{2}$$

i.e. $\mathrm{SD} \cup H \cup O$ *is consistent.*

Here, $\nvDash$ stands for the negation of the logical entailment relation, and $\bot$ represents 'falsum'. The consistency condition (2) captures DNSB diagnosis in terms of consistency-based diagnosis under the assumption that the axioms in SD provide a completely accurate and correct representation of a physical system. A diagnosis is just a hypothesis that is accepted. In the formalisation by De Kleer et al., [De Kleer et al., 1992], each literal $Abnormal(c) \in H$

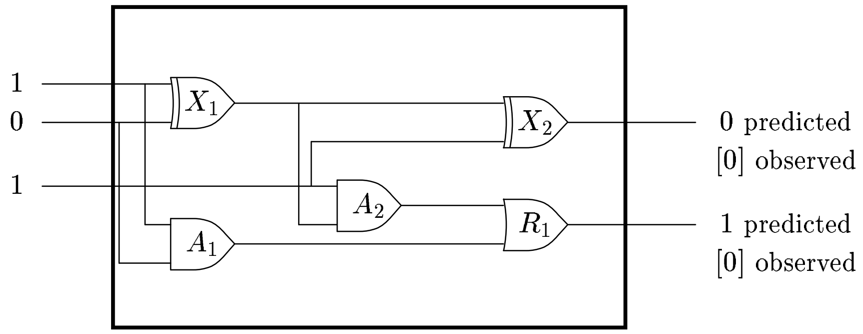

Figure 5: Full adder.

is interpreted as being defective; a literal $\neg Abnormal(c) \in H$ indicates component c to be nondefective. In the original theory by Reiter, [Reiter, 1987], the set D above is taken as a diagnosis, with the extra requirement that D is minimal with respect to set inclusion. Then, each component c in a diagnosis D for which $Abnormal(c)$ is *true* is interpreted as being defective. According to expression (2), taking $D =$ COMPS leads to the trivial diagnosis that all components are defective (or the defective components are among the set of all components). Reiter, therefore, incorporated in the original theory the requirement that the set D must be a minimal set with respect to set inclusion, fulfilling the consistency condition. However, later it was recognised that minimality according to set inclusion is merely a measure of plausibility, which may not be appropriate when knowledge of abnormal behaviour is also included in the system description SD, and the minimality criterion was left out of the basic definition [De Kleer et al., 1992]. Moreover, other measures of plausibility (cf. [Tuhrim et al., 1991] in the context of abduction) may also apply. The application of the formal theory by Reiter is illustrated by a classical example from the literature [Genesereth, 1984].

Example 1. Consider the logical circuit depicted in Figure 5, which represents a full adder, i.e. a circuit that can be used for the addition of two bits with carry-in and carry-out bits. The components X_1 and X_2 represent exclusive-OR gates, A_1 and A_2 represent AND gates, and R_1 represents an OR gate.

The system description, as provided in [Reiter, 1987], consists of the following axioms:

$$\begin{aligned}
\forall x(\mathrm{ANDG}(x) \wedge \neg Abnormal(x) &\rightarrow out(x) = and(in1(x), in2(x))) \\
\forall x(\mathrm{XORG}(x) \wedge \neg Abnormal(x) &\rightarrow out(x) = xor(in1(x), in2(x))) \\
\forall x(\mathrm{ORG}(x) \wedge \neg Abnormal(x) &\rightarrow out(x) = or(in1(x), in2(x)))
\end{aligned}$$

which describe the (normal) behaviour of each individual component (gate), and

$$\begin{aligned}
out(X_1) &= in2(A_2) \\
out(X_1) &= in1(X_2) \\
out(A_2) &= in1(R_1) \\
in1(A_2) &= in2(X_2) \\
in1(X_1) &= in1(A_1) \\
in2(X_1) &= in2(A_1) \\
out(A_1) &= in2(R_1)
\end{aligned}$$

which gives information about the connections between the components, i.e. information about the normal structure, including some electrical relationships. Finally, the various gates are defined:

$$\begin{array}{l} \mathrm{ANDG}(A_1) \\ \mathrm{ANDG}(A_2) \\ \mathrm{XORG}(X_1) \\ \mathrm{XORG}(X_2) \\ \mathrm{ORG}(R_1) \end{array}$$

Appropriate axioms for a Boolean algebra are also assumed to be available.

Now, let us assume that

$$O = \{\mathit{in1}(X_1) = 1, \mathit{in2}(X_1) = 0, \mathit{in1}(A_2) = 1, \mathit{out}(X_2) = 0, \mathit{out}(R_1) = 0\}$$

Note that $\mathit{out}(R_1) = 1$ is predicted using the model of normal structure and behaviour in Figure 5, which is in contrast with the observed output $\mathit{out}(R_1) = 0$. Assuming that $H = \{\neg \mathit{Abnormal}(c) \mid c \in \mathrm{COMPS}\}$, it follows that

$$\mathrm{SD} \cup H \cup O$$

is inconsistent. This confirms that some of the output signals observed differ from those expected under the assumption that the circuit is functioning normally. Using Formula (2), a possible diagnosis is, for instance,

$$\begin{array}{ll} H' = \{\mathit{Abnormal}(X_1), \neg\mathit{Abnormal}(X_2), \neg\mathit{Abnormal}(A_1), \\ \qquad \neg\mathit{Abnormal}(A_2), \neg\mathit{Abnormal}(R_1)\} \end{array}$$

since

$$\mathrm{SD} \cup H' \cup O$$

is consistent. In terms of Reiter's original definition, the corresponding diagnosis would be $D' = \{X_1\}$. Note that, given the diagnosis H', no output is predicted for the circuit; the assumption $\mathit{Abnormal}(X_1)$ completely blocks transforming input into output by the modelled circuit, because

$$\mathrm{SD} \cup H' \cup O \backslash \{\mathit{out}(X_2) = 0\} \nvDash \mathit{out}(X_2) = 0$$

In a sense, this is too much, because there was no discrepancy between the predicted and observed output of gate X_2. Nevertheless, the hypothesis H' is a diagnosis according to Definition 1. ◇

It is interesting to look at consistency-based diagnosis in a more intuitive way. What the theory actually expresses is that if components that may be defective are removed from a system or device, and the resulting newly predicted behaviour, or no behaviour at all, does not contradict the observed behaviour, then a diagnosis has been established. This is a rather crude approach to diagnosis. Imagine that we have a formal model of an electrical device, including its electric plug, then simulating the removal of the plug from its socket, thus recovering consistency, will provide us with a diagnosis for a defective system. According to the theory, the plug will be identified as the culprit, which, of course, is absurd if the device was in operation prior to the removal of the plug, although incorrectly. K. Konolige, [Konolige, 1994], refers to diagnoses produced by consistency-based diagnosis as *excuses*, to reflect that it may not be possible to explain such diagnoses in terms of cause-effect relationships.

In addition to a definition of consistency-based diagnosis, [De Kleer et al., 1992] introduces the concepts of partial diagnosis and kernel diagnosis. A *partial diagnosis* is an abbreviated representation for a set of diagnoses that have certain '*Abnormal*' and '$\neg$*Abnormal*' literals in common. For example, in addition to H' in Example 1,

$$\begin{aligned} H'' = \{&\mathit{Abnormal}(X_1), \mathit{Abnormal}(X_2), \neg\mathit{Abnormal}(A_1),\\ &\neg\mathit{Abnormal}(A_2), \neg\mathit{Abnormal}(R_1)\} \end{aligned}$$

is also a diagnosis. The two diagnoses H' and H'' can be abbreviated as the partial diagnosis

$$P = \{\mathit{Abnormal}(X_1), \neg\mathit{Abnormal}(A_1), \neg\mathit{Abnormal}(A_2), \neg\mathit{Abnormal}(R_1)\}$$

which explicitly indicates that the actual status of component X_2 is irrelevant, adopting for the other components the status mentioned in the partial diagnosis P. Note that a partial diagnosis is not a real diagnosis according to Definition 1, because not all components are assigned unique '*Abnormal*' modes. A *kernel diagnosis* is simply a partial diagnosis that is minimal with respect to set inclusion.

Above, it was assumed that a system description SD is expressed using standard logic, using standard, monotonic logical entailment to define the notion of consistency-based diagnosis, but this is not essential. We may as well use some nonmonotonic logic. However, even when restricted to standard, monotonic logic, the notion of consistency-based diagnosis is nonmonotonic: observing additional findings may result in cancelling prior diagnoses [Reiter, 1987].

Example 2. Reconsider the system description SD from Example 1. Assume that

$$O = \{\mathit{in1}(X_1) = 1, \mathit{in2}(X_1) = 0, \mathit{in1}(A_2) = 1, \mathit{out}(X_2) = 0\}$$

Then, the diagnosis is equal to

$$\begin{aligned} H'' = \{&\neg\mathit{Abnormal}(X_1), \neg\mathit{Abnormal}(X_2), \neg\mathit{Abnormal}(A_1),\\ &\neg\mathit{Abnormal}(A_2), \neg\mathit{Abnormal}(R_1)\} \end{aligned}$$

or, in Reiter's original notation: $D'' = \varnothing$ (there are no faults). Observing $\mathit{out}(R_1) = 0$ yields, among others, the diagnosis mentioned in Example 1 ($D' = \{X_1\}$ in Reiter's notation), but H'' ($D'' = \varnothing$) is not longer a diagnosis. $\diamond$

Reiter, [Reiter, 1987], has also given an analysis of consistency-based diagnosis in terms of default logic (cf. [Besnard, 1989]). A system description SD and a set of observed findings O are supplemented with default rules of the form

$$\frac{: \neg\mathit{Abnormal}(c)}{\neg\mathit{Abnormal}(c)}$$

for each component c, yielding a default theory. A default rule as above expresses that $\neg\mathit{Abnormal}(c)$ may be assumed for component c, if assuming $\neg\mathit{Abnormal}(c)$ does not give rise to inconsistency. Hence, in computing an extension of the resulting default theory, these default rules will only be applied under the condition that they do not violate consistency, which is precisely the effect of the consistency condition (2). This mapping of a system $\mathcal{S}$ to default logic offers an object-level characterisation of the meta-level description of consistency-based diagnosis given in Definition 1.

In Section 5, the application of Reiter's theory to the logical formalisation of MAB diagnosis will be discussed. The techniques proposed by Reiter are not the only possible ways to formalise DNSB and MAB diagnosis; D. Poole has proposed other logical techniques for the same purpose in terms of his Theorist framework of hypothetical reasoning [Poole et al., 1987; Poole, 1990a; Poole, 1990b; Poole, 1994]. This work, however, bears great resemblance to the work by Reiter with respect to DNSB diagnosis, and to the work by Console and Torasso with respect to MAB diagnosis, which will be discussed in the following section. The Theorist framework is discussed in Section 5.

4.2 Abductive diagnosis

The formalisation of MAB diagnosis has been extensively studied by L. Console and P. Torasso [Console et al., 1989; Console & Torasso, 1990a]. In their theory, the abnormal behaviour of a system is specified in terms of abnormal states and resulting abnormal findings. Normal findings may also be included, but these are less useful for diagnosis, since an abnormal state is often causally related to a large number of normal findings. Diagnostic problem solving is formally described as the problem of accounting for a given set of observed findings, referred to in the theory as manifestations, by the simulation of abnormal behaviour. The simulation process is accomplished by deduction with logical axioms, describing abnormal behaviour, and assumed (abnormal) states.

The logical axioms are Horn clauses of the following form and meaning

$$\begin{aligned} State_1 \wedge \cdots \wedge State_n &\rightarrow o & (3)\\ State_1 \wedge \cdots \wedge State_n &\rightarrow State & (4)\\ State_1 \wedge \cdots \wedge State_n &\rightarrow d & (5)\end{aligned}$$

where $State$ and $State_i$, $i = 1, \ldots, n$, are positive literals representing part of the internal state of a modelled system, d is a *defect* (or disorder), and o is an *observable finding.* In a number of articles, extension to general Horn clauses (Horn clauses with negation as failure) is proposed (cf. [Console et al., 1991; Preist et al., 1994]). For simplicity's sake, we shall adopt the Horn-clause restriction in this section. It is assumed that the set of Horn clauses is hierarchical, i.e. no cyclic dependencies among atoms in clauses are allowed (which contrasts with the situation in logic programming, where cyclic dependencies are almost the rule). In the original abductive theory of diagnosis by Console and Torasso, as described in [Console et al., 1989], a finding appearing in the conclusion of a logical implication represents an *abnormal* finding. In [Console & Torasso, 1990b], however, *normal* findings are also allowed. Recall that, when necessary for clarity, abnormal findings are denoted by o_{ab}; similarly, normal findings are denoted by o_{norm}.

A state literal is employed for the simulation of the occurrence of (abnormal) behaviour using the logical specification. It corresponds to a parameter with a value. For example, if the parameter *pressure*(*blood*) can take values *decreased*, *normal* and *increased*, then

$$pressure(blood) = increased$$

corresponds to a state. The intuitive meaning of formulae of the form (3) is: 'presence of $State_1, \ldots, State_n$ *causes* the (abnormal) finding o', i.e. if $State_1, \ldots, State_n$ hold in the system, (abnormal) finding o must be observed. Formulae of the form (4) express that a collection of states is causally related to another state, i.e. if the states $State_1, \ldots, State_n$ occur then *State* occurs as well. Axioms that conform to the two axiom schemata above are sometimes referred to as *abnormality axioms*. Note that the notion of causality is expressed in the theory using logical implication. Logical implication is employed to express a causal relationship between states and observable findings, and between states and states. Axioms of the form (5) can be viewed as *classification axioms* because they classify a collection of states as a particular defect. The idea originates from the CASNET system [Weiss et al., 1978]. If sufficient state literals are assumed or derived to satisfy the antecedent of an axiom of the form (5), a defect d can be derived. In the theory by Console and Torasso, a defect is actually defined in terms of a collection of states. This can be expressed by using a bi-implication ($\leftrightarrow$) instead of an implication, as in axiom schema (5). However, when adopting this formalisation for diagnosis, the implications from right to left ($\leftarrow$) are not involved. Classification axioms are not an essential ingredient of the theory of diagnosis by Console and Torasso; they are merely used to attach diagnostic labels to collections of states. Note that in the classification axioms, logical implication is used to express a classification instead of a causal relationship, as in the abnormality axioms. Due to the manifold uses of logical implication, the theory provides no clear logical meaning for the various relationships, including causality, underlying the theory of diagnosis. To express the theory in terms of defects and findings only, thus enabling us to analyse the essentials of the theory, states are identified with defects. Thus, axioms of the form (4) and (5) are collapsed into one axiom schema; the classification axioms are given no further consideration. In the following, it shall be assumed that axioms are of the following two forms:

$$d_1 \wedge \cdots \wedge d_n \quad \rightarrow \quad o \tag{6}$$

$$d_1 \wedge \cdots \wedge d_n \quad \rightarrow \quad d \tag{7}$$

where d, d_i, $i = 1, \ldots, n$, represent defects. We shall try to convey the essentials of the theory, using the uniform terminology and notation adopted in this article, thus deviating in some respects from the original papers.

Console and Torasso also provide a mechanism in their logical formalisation to weaken the causality relation. To this end, literals α are introduced into the premises of the axioms of the form (6) and (7), which can be used to block the deduction of an observable finding o or defect d if the defects d_i, $i = 1, \ldots, n$, hold true, by assuming the literal α to be false. The weakened axioms have the following form:

$$d_1 \wedge \cdots \wedge d_n \wedge \alpha_o \quad \rightarrow \quad o \tag{8}$$

$$d_1 \wedge \cdots \wedge d_n \wedge \alpha_d \quad \rightarrow \quad d \tag{9}$$

The literals α are called *incompleteness-assumption literals*, abbreviated to *assumption literals*. Axioms of the form (6) – (9) are now taken as the (abnormality) axioms.

In the following, let $\mathcal{C} = (\mathrm{DFS}, \mathrm{OBS}, \mathrm{CM})$ stand for a *causal specification* in the theory of diagnosis by Console and Torasso, where:

- DFS denotes a set of possible defect and assumption literals;
- OBS denotes a set of possible (positive and negative) observable finding literals;
- CM ('Causal Model') stands for a set of logical (abnormality) axioms of the form (6) – (9).

Subsets of the set DFS will be called *hypotheses*. A causal specification can then be employed for the prediction of observable findings in the sense of Figure 3.

Definition 2 (*prediction*). *Let* $\mathcal{C} = (\mathrm{DFS}, \mathrm{OBS}, \mathrm{CM})$ *be a causal specification. Then, a hypothesis* $H \subseteq \mathrm{DFS}$ *is called a* prediction *for a set of observable findings* $O \subseteq \mathrm{OBS}$ *if*

(1) $\mathrm{CM} \cup H \vDash O$*, and*

(2) $\mathrm{CM} \cup H$ *is consistent.*

Hence, the notion of prediction formalises the arrow in the lower half of Figure 3; the resulting set of findings O corresponds to the predicted (observable) findings in the same figure.

An *abductive diagnostic problem* $\mathcal{A}$ is now defined as a pair $\mathcal{A} = (\mathcal{C}, O)$, where $O \subseteq \mathrm{OBS}$ is called a *set of observed findings*. A set of observed findings corresponds to the box in the upper half of Figure 3.

Formally, a solution to an abductive diagnostic problem $\mathcal{A}$ can be defined as follows.

Definition 3 (*solution*). *Let* $\mathcal{A} = (\mathcal{C}, O)$ *be an abductive diagnostic problem, where* $\mathcal{C} = (\mathrm{DFS}, \mathrm{OBS}, \mathrm{CM})$ *is a causal specification with* CM *a set of abnormality axioms of the form (6) – (9), and* $O \subseteq \mathrm{OBS}$ *a set of observed findings. A hypothesis* $H \subseteq \mathrm{DFS}$ *is called a* solution *to* $\mathcal{A}$ *if:*

(1) $\forall o \in O : \ \mathrm{CM} \cup H \vDash o$ (covering condition)*;*

(2) $\forall o \in O^c : \ \mathrm{CM} \cup H \nvDash \neg o$ (consistency condition)

where O^c *is defined by:*

$$O^c = \{\neg o \in \mathrm{OBS} \mid o \in \mathrm{OBS}, o \notin O, \ o \textit{ is a positive literal}\}$$

In the work by Console and Torasso, the set $\mathrm{CM} \cup H$ is called a 'world' if H is a prediction; the set $\mathrm{CM} \cup H$ is called a 'final world' if H is a solution to an abductive diagnostic problem [Console et al., 1989; Console & Torasso, 1990a]. Note that the sets O and O^c are disjoint, and that if $o \in O$ then $\neg o \notin O^c$. The set O^c stands for findings assumed to be false, because they have not been observed (and are therefore assumed to be absent). But any finding may also be unknown. Thus, rather than providing a single definition, Console and Torasso provide in their articles several alternatives for this set O^c. The definition provided in Definition 3 above is just one of the alternatives.

Condition (1) is called the covering condition, because it requires that each observed finding is accounted for by a solution H. Note that any solution to a diagnostic problem $\mathcal{A} = (\mathcal{C}, O)$ is a prediction for O according to Definition 2. Condition (2) is called the consistency condition, because it can be restated as follows

$$\mathrm{CM} \cup H \cup O^c \nvDash \bot$$

A set of defects in a prediction H is also called a set of *perturbations* [Console & Torasso, 1990a]; in [Console & Torasso, 1991] the term *abducibles* is employed for literals that may be assumed as part of diagnostic problem solving.

In the original formulation of the theory only those defects (states) are admitted to H which do not appear in the conclusions of implications; such defects are called *initial defects* (initial states in the original theory). The covering condition defined above ensures that sufficient defects and assumption literals are assumed to account for all given observed findings. The consistency condition helps to ensure that not too many defect and assumption literals are assumed. Although it is only necessary to include an assumption literal α in a solution for implications $d \wedge \alpha_o \to o$ and $d \wedge \alpha_{d'} \to d'$ if the defect d is deducible from the assumed (initial) defects and assumption literals, Definition 3 does not always prevent their inclusion in a solution.

An entire solution H may be taken as a diagnosis, but following [Console et al., 1989], a diagnosis is considered to consist of the defect literals in a solution H.

Definition 4 (*abductive diagnosis*). *Let $\mathcal{A} = (\mathcal{C}, O)$ be an abductive diagnostic problem, where $\mathcal{C} = (\mathrm{DFS}, \mathrm{OBS}, \mathrm{CM})$ is a causal specification. Let H be a solution to $\mathcal{A}$. Then, the set of all defects $D \subseteq H$ is called an* (abductive) diagnosis *of $\mathcal{A}$.*

Recall that in [Console et al., 1989], a diagnosis is obtained by applying the classification axioms (5); a distinction is therefore made in [Console et al., 1989] between a solution H for which the covering and consistency conditions are satisfied, i.e. the set of defect and assumption literals contained in a 'final world' – this world is called a *causal explanation* – and the set of defects resulting from an explanation, which is called a diagnosis (originally, a solution). However, from a formal point of view, the distinction is not essential.

Example 3. Consider the causal specification $\mathcal{C} = (\mathrm{DFS}, \mathrm{OBS}, \mathrm{CM})$, with

$$\mathrm{DFS} = \{\mathit{fever}, \mathit{influenza}, \mathit{sport}, \alpha_1, \alpha_2\}$$

and

$$\mathrm{OBS} = \{\mathit{chills}, \mathit{thirst}, \mathit{myalgia}, \neg\mathit{chills}, \neg\mathit{thirst}, \neg\mathit{myalgia}\}$$

'Myalgia' means painful muscles. The following set of logical formulae CM, representing medical knowledge concerning influenza and sport, both 'disorders' with frequent occurrence, is given:

$$\begin{array}{l} \mathit{fever} \wedge \alpha_1 \to \mathit{chills} \\ \mathit{influenza} \to \mathit{fever} \\ \mathit{fever} \to \mathit{thirst} \\ \mathit{influenza} \wedge \alpha_2 \to \mathit{myalgia} \\ \mathit{sport} \to \mathit{myalgia} \end{array}$$

For example, $\mathit{influenza} \wedge \alpha_2 \to \mathit{myalgia}$ means that influenza *may cause* myalgia; $\mathit{influenza} \to \mathit{fever}$ means that influenza *always causes* fever. For illustrative purposes, a causal knowledge base as given above is often depicted as a labelled, directed graph G, which is called a *causal net*, as shown in Figure 6. Suppose that the abductive diagnostic problem $\mathcal{A} = (\mathcal{C}, O)$ must be solved, where the set of observed findings $O = \{\mathit{thirst}, \mathit{myalgia}\}$. Then, $O^c = \{\neg\mathit{chills}\}$. There are several solutions to this abductive diagnostic problem (for which the consistency and covering conditions are fulfilled):

$$\begin{array}{l} H_1 = \{\mathit{influenza}, \alpha_2\} \\ H_2 = \{\mathit{influenza}, \mathit{sport}\} \\ H_3 = \{\mathit{fever}, \mathit{sport}\} \end{array}$$

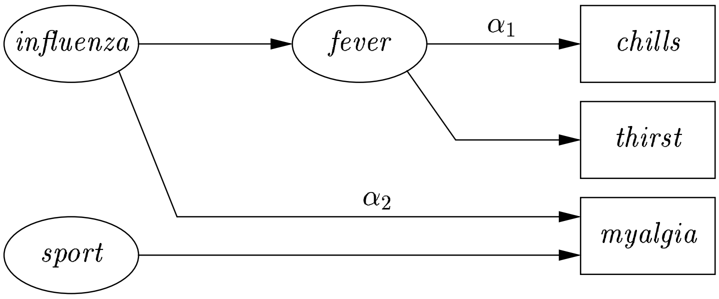

Figure 6: A knowledge base with causal relations.

$$
\begin{aligned}
H_4 &= \{\mathit{fever}, \mathit{influenza}, \alpha_2\}\\
H_5 &= \{\mathit{influenza}, \alpha_2, \mathit{sport}\}\\
H_6 &= \{\mathit{fever}, \mathit{influenza}, \mathit{sport}\}\\
H_7 &= \{\mathit{fever}, \mathit{influenza}, \alpha_2, \mathit{sport}\}
\end{aligned}
$$

The following diagnoses correspond to these solutions:

$$
\begin{aligned}
D_1 &= \{\mathit{influenza}\}\\
D_2 &= \{\mathit{influenza}, \mathit{sport}\}\\
D_3 &= \{\mathit{fever}, \mathit{sport}\}\\
D_4 &= \{\mathit{fever}, \mathit{influenza}\}\\
D_5 &= \{\mathit{fever}, \mathit{influenza}, \mathit{sport}\}
\end{aligned}
$$

For example, the diagnosis $D_4 = \{\mathit{fever}, \mathit{influenza}\}$ means that the patient has influenza with associated fever. Restricting to initial defects would yield the solutions H_1, H_2 and H_5 and the diagnoses D_1 and D_2. Finally, note that, for example, the hypothesis $H = \{\alpha_1, \alpha_2, \mathit{fever}, \mathit{influenza}\}$ is incompatible with the consistency condition. ◇

Because in this theory of diagnosis, the observable findings are logically entailed by the assumption of the presence of certain states, and the reasoning goes in a sense in a direction reverse to that of the logical implication, i.e. from the consequent to the premise, the theory is often referred to as the *abductive theory of diagnosis*, or *abductive diagnosis* for short.

Several researchers (cf. [Poole, 1988; Console et al., 1991]) have noted a close correspondence between abduction and the predicate completion of a logical theory, as originally proposed by K. Clark in connection with negation as finite failure in logic programming [Clark, 1978]. Consider the following example.

Example 4. Suppose that sport and influenza are two 'disorders'; this may be expressed in predicate logic as follows:

$$
\begin{aligned}
&\mathit{Disorder}(\mathit{sport})\\
&\mathit{Disorder}(\mathit{influenza})
\end{aligned}
$$

The following logical implication is equivalent to the conjunction of the two literals above:

$$\forall x((x = \mathit{sport} \vee x = \mathit{influenza}) \to \mathit{Disorder}(x))$$

assuming the presence of the logical axioms for equality, and also assuming that constants with different names are not equal. Suppose that sport and influenza are the *only* possible disorders. This can be expressed by adding the following logical implication:

$$\forall x(\mathit{Disorder}(x) \rightarrow (x = \mathit{sport} \vee x = \mathit{influenza})) \tag{10}$$

to the implication above. For example, adding *Disorder*(*asthma*) to logical implication (10) yields an inconsistency, because *asthma* is neither equal to *sport* nor equal to *influenza*: the conclusion

$$\mathit{asthma} = \mathit{sport} \vee \mathit{asthma} = \mathit{influenza}$$

cannot be satisfied. Now, suppose that the literal *Disorder*(*asthma*) is removed, but that '*asthma*' remains a valid constant symbol. Then, $\neg$*Disorder*(*asthma*) is a logical consequence of formula (10); this formula 'completes' the logical theory by stating that disorders not explicitly mentioned are assumed to be false. Formula (10) is called a *completion formula.*$\diamond$

The characterisation of abduction as deduction in a completed logical theory is natural, because computation of the predicate completion of a logical theory amounts to adding the only-if parts of the formulae to the theory, i.e. it 'reverses the arrow' which is exactly what happens when abduction is applied to derive conclusions. After all, abductive reasoning is reasoning in a direction reverse to logical implication. In an intuitive sense, predicate completion expresses that the only possible causes (defects) for observed findings are those appearing in the abnormality axioms; assumption literals are taken as implicit causes. Where the characterisation of abduction by means of the covering and consistency conditions may be viewed as a meta-level description of abductive diagnosis, the predicate completion can be taken as the object-level characterisation, i.e. in terms of the original axioms in CM. [Poole, 1988] and [Console et al., 1991] note that, in contrast to the predicate completion in logic programming, predicate completion should only pertain to literals appearing as a consequence of the logical axioms in CM, i.e. finding literals and defect literals that can be derived from other defects and assumption literals. This set of defects and observable findings is called the set of *non-abducible* literals, denoted by A; the set DFS$\backslash A$ is then called the set of *abducible* literals.

Let us denote the axiom set CM by

$$\begin{aligned}\mathrm{CM} = \{&\varphi_{1,1} \rightarrow a_1, \ldots, \varphi_{1,n_1} \rightarrow a_1,\\ &\vdots\\ &\varphi_{m,1} \rightarrow a_m, \ldots, \varphi_{m,n_m} \rightarrow a_m\}\end{aligned}$$

where $A = \{a_i \mid 1 \leq i \leq m\}$ is the set of non-abducible (finding or defect) literals and each $\varphi_{i,j}$ denotes a conjunction of defect literals, possibly including an assumption literal. The predicate completion of CM with respect to the non-abducible literals A, denoted by COMP[CM; A] (cf. [Genesereth & Nilsson, 1987]), is defined as follows:

$$\begin{aligned}\mathrm{COMP}[\mathrm{CM}; A] = \mathrm{CM} \cup \{&a_1 \rightarrow \varphi_{1,1} \vee \cdots \vee \varphi_{1,n_1},\\ &\vdots\\ &a_m \rightarrow \varphi_{m,1} \vee \cdots \vee \varphi_{m,n_m}\}\end{aligned}$$

The predicate completion of CM makes explicit the fact that the only causes of non-abducible literals (findings and possibly also defects) are the defects and assumption literals given as a disjunct in the consequent. For example,

$$o_{ab} \rightarrow d_1 \vee \cdots \vee d_n$$

indicates that only the defects from the set $\{d_1, \ldots, d_n\}$ can be used to explain the observed finding o_{ab}.

Predicate completion of abnormality axioms with respect to a set of non-abducible literals can now be used to characterise diagnosis. Let ψ and ψ' be two logical formulae. It is said that ψ is *more specific than* ψ' iff $\psi \vDash \psi'$. Using the predicate completion of a set of abnormality axioms CM, we now have the following definition.

Definition 5 (*solution formula*)**.** *Let $\mathcal{A} = (\mathcal{C}, O)$ be an abductive diagnostic problem and let* $\mathrm{COMP}[\mathrm{CM}; A]$ *be the predicate completion of* CM *with respect to A, the set of non-abducible literals in $\mathcal{A}$. A* solution formula *S for $\mathcal{A}$ is defined as the most specific formula consisting only of abducible literals, such that*

$$\mathrm{COMP}[\mathrm{CM}; A] \cup O \cup O^c \vDash S$$

where O^c is defined as in Definition 3.

Hence, abductive diagnosis is transformed to hypothetico-deductive diagnosis (cf. Section 4.4). A solution formula is obtained by applying the set of equivalences in $\mathrm{COMP}[\mathrm{CM}; A]$ to a set of observed findings O, augmented with those findings not observed, O^c, yielding a logical formula that includes all possible solutions according Definition 3, given the equivalences in $\mathrm{COMP}[\mathrm{CM}; A]$. The following theorem, which is proven in [Console et al., 1991], reveals an important relationship between the meta-level characterisation of abductive diagnosis, as presented in Definition 3, and the object-level characterisation of diagnosis in Definition 5.[2]

Theorem 1. *Let $\mathcal{A} = (\mathcal{C}, O)$ be an abductive diagnostic problem, where $\mathcal{C} = (\mathrm{DFS}, \mathrm{OBS}, \mathrm{CM})$ is a causal specification. Let O^c be defined as in Definition 3, and let S be a solution formula for $\mathcal{A}$. Let $H \subseteq \mathrm{DFS}$ be a set of abducible literals, and let I be an interpretation of $\mathcal{A}$, such that for each abducible literal $a \in \mathrm{DFS}$: $\vDash_I a$ iff $a \in H$. Then, H is a solution to $\mathcal{A}$ iff $\vDash_I S$.*

Proof. ($\Rightarrow$): The set of defect and assumption literals H is a solution to $\mathcal{A}$, hence, for each $o \in O$: $\mathrm{CM} \cup H \vDash o$, and for each $o' \in O^c$: $\mathrm{CM} \cup H \nvDash \neg o'$. The solution formula S is the result of rewriting observed findings in O and non-observed findings in O^c using the equivalences in $\mathrm{COMP}[\mathrm{CM}; A]$ to a formula merely consisting of abducibles. Assume that S is in conjunctive normal form. Conjuncts in S are equivalent to observed findings $o \in O$, that are logically entailed by $\mathrm{CM} \cup H$, or to non-observed findings $\neg o \in O^c$ that are consistent with $\mathrm{CM} \cup H$. Hence, an interpretation I for which $\vDash_I H$, that falsifies each abducible in $\mathrm{DFS} \backslash H$, satisfying every $o \in O$ and each $\neg o \in O^c$ that has been rewritten, must satisfy this collection of conjuncts, i.e. S.
($\Leftarrow$): If S is in conjunctive normal form, S must be the result of rewriting observed findings $o \in O$ and non-observed findings in O^c to (negative or positive) abducibles, using the equivalences in $\mathrm{COMP}[\mathrm{CM}; A]$. Since an interpretation I that satisfies H and S must also satisfy each finding $o \in O$ and those $\neg o \in O^c$ that have been rewritten to S, it follows that I can be chosen such that $\vDash_I O^c$, i.e. H must be a solution to $\mathcal{A}$. ◇

This theorem reveals an important property of the abductive theory of diagnosis. Sometimes, a solution to an abductive diagnostic problem is capable of satisfying a solution formula in

[2] Contrary to our treatment, in [Console et al., 1991], a solution H of an abductive problem $\mathcal{A}$ is defined by SLD resolution with the negation as finite failure rule, i.e. SLDNF resolution, such that $\mathrm{CM} \cup H \vdash_{\mathrm{SLDNF}} O \cup O^c$, i.e. the covering and consistency conditions are merged.

the technical, logical sense.

Example 5. Reconsider the set of logical axioms given in Example 3. The predicate completion of CM is equal to

$$\begin{aligned}
&\mathrm{COMP}[\mathrm{CM}; \{\mathit{chills}, \mathit{thirst}, \mathit{myalgia}, \mathit{fever}\}]\\
= {} & \mathrm{CM} \cup \{\mathit{chills} \rightarrow \mathit{fever} \wedge \alpha_1,\\
&\qquad \mathit{fever} \rightarrow \mathit{influenza},\\
&\qquad \mathit{thirst} \rightarrow \mathit{fever},\\
&\qquad \mathit{myalgia} \rightarrow (\mathit{influenza} \wedge \alpha_2) \vee \mathit{sport}\}\\
= {} & \{\mathit{chills} \leftrightarrow \mathit{fever} \wedge \alpha_1,\\
&\ \mathit{fever} \leftrightarrow \mathit{influenza},\\
&\ \mathit{thirst} \leftrightarrow \mathit{fever},\\
&\ \mathit{myalgia} \leftrightarrow (\mathit{influenza} \wedge \alpha_2) \vee \mathit{sport}\}
\end{aligned}$$

Note that

$$\begin{aligned}
\mathrm{COMP}[\mathrm{CM}; \{\mathit{chills}, \mathit{thirst}, \mathit{myalgia}, \mathit{fever}\}] \cup O \cup O^c \vDash\\
(\mathit{influenza} \wedge \alpha_2) \vee (\mathit{influenza} \wedge \mathit{sport})
\end{aligned}$$

given that $O = \{\mathit{thirst}, \mathit{myalgia}\}$ and $O^c = \{\neg\mathit{chills}\}$. Although

$$\mathrm{COMP}[\mathrm{CM}; \{\mathit{chills}, \mathit{thirst}, \mathit{myalgia}, \mathit{fever}\}] \cup O \cup O^c \vDash \neg(\mathit{fever} \wedge \alpha_1)$$

the formula $\neg(\mathit{fever} \wedge \alpha_1)$, which is a logical consequence of $\neg\mathit{chills}$ and $\mathit{chills} \leftrightarrow (\mathit{fever} \wedge \alpha_1)$, is not part of the solution formula $S \equiv (\mathit{influenza} \wedge \alpha_2) \vee (\mathit{influenza} \wedge \mathit{sport})$, because the literal *fever* is non-abducible. It holds, in accordance with Theorem 1, that

$$\vDash_I H_i \ \Rightarrow\ \vDash_I (\mathit{influenza} \wedge \alpha_2) \vee (\mathit{influenza} \wedge \mathit{sport})$$

for $i = 1, 2, 5$, where H_i is a solution given in Example 3 consisting only of abducible literals, for suitable interpretations I. Here, it even holds that $H_i \vDash S$, because S does not contain any negative defects or assumption literals entailed by non-observed findings in O^c. ◇

Although the theory by Console and Torasso is restricted to reasoning with causal domain knowledge, other types of knowledge, referred to as *contextual information* by Console and Torasso, is also dealt with in the theory. Contextual information is incorporated to render the causal relation conditional on certain findings, e.g. in

$$d \wedge o \rightarrow o'$$

the finding literal o acts as a condition with regard to the causal relation between the defect d and the finding o'. For example, in a medical setting, many causal relations are age-specific; hence, the observed (normal) finding '$\mathit{age} \circ v$', where $\circ$ denotes an ordering predicate and v an integer, could be employed to express such conditional causality.

Above we have defined abductive diagnosis using propositional logic. The definition in terms of predicate logic reveals some additional subtleties, yielding various alternative definition for the set of findings not observed and assumed to be absent, O^c. Findings o are denoted in predicate logic using a predicate symbol p, indicating a particular group of findings or a test. For example, in '$\mathit{Sign}(\mathit{fever})$', the predicate symbol '*Sign*' denotes a group of patient findings; in '$\mathit{Serum\_copper}(\mathit{patient}, \mathit{high})$', the predicate symbol '$\mathit{Serum\_copper}$' indicates the result of a diagnostic test. The consequences of using predicate logic to define abductive diagnosis will be briefly introduced by means of the following example.

Example 6. Consider the following (partial) set of abnormality axioms CM, expressed in first-order predicate logic as follows:

$$\begin{aligned}
\mathit{Disorder}(\mathit{influenza}) &\rightarrow \mathit{Symptom}(\mathit{cough})\\
\mathit{Disorder}(\mathit{influenza}) &\rightarrow \mathit{Sign}(\mathit{fever})\\
\mathit{Disorder}(\mathit{pulmonary\_embolism}) &\rightarrow \mathit{Blood\_chemistry}(O_2\text{-}\mathit{level}, \mathit{low})
\end{aligned}$$

where the (ground) literals $\mathit{Symptom}(\mathit{cough})$, $\mathit{Sign}(\mathit{fever})$ and $\mathit{Blood\_chemistry}(O_2\text{-}\mathit{level}, \mathit{low})$ stand for observable findings, and the '$\mathit{Disorder}$' literals represent defects. The finding literals o, representing abnormal observable findings, are taken from the following set of positive finding literals:

$$\begin{aligned}
\mathrm{OBS}_P = \{&\mathit{Symptom}(\mathit{cough}), \mathit{Symptom}(\mathit{headache}),\\
&\mathit{Sign}(\mathit{fever}), \mathit{Sign}(\mathit{hypertension}),\\
&\mathit{Blood\_chemistry}(O_2\text{-}\mathit{level}, \mathit{low}), \mathit{Blood\_chemistry}(\mathit{Sodium}, \mathit{low})\}
\end{aligned}$$

and the set of negative finding literals is equal to

$$\begin{aligned}
\mathrm{OBS}_N = \{&\neg\mathit{Symptom}(\mathit{cough}), \neg\mathit{Symptom}(\mathit{headache}),\\
&\neg\mathit{Sign}(\mathit{fever}), \neg\mathit{Sign}(\mathit{hypertension}),\\
&\neg\mathit{Blood\_chemistry}(O_2\text{-}\mathit{level}, \mathit{low}), \neg\mathit{Blood\_chemistry}(\mathit{Sodium}, \mathit{low})\}
\end{aligned}$$

with $\mathrm{OBS} = \mathrm{OBS}_P \cup \mathrm{OBS}_N$. Now, let $O = \{\mathit{Sign}(\mathit{fever}), \mathit{Blood\_chemistry}(O_2\text{-}\mathit{level}, \mathit{low})\}$ be a set of observed findings. Usually, it is assumed that $O \subseteq \mathrm{OBS}_P$, because only positive findings can be accounted for by Horn clauses in CM. The set $O^c \subseteq \mathrm{OBS}$, representing the findings not observed, is constructed in accordance with Definition 3. In the present case, the set O^c is equal to

$$\begin{aligned}
O^c = \{&\neg\mathit{Symptom}(\mathit{cough}), \neg\mathit{Symptom}(\mathit{headache}),\\
&\neg\mathit{Sign}(\mathit{hypertension}),\\
&\neg\mathit{Blood\_chemistry}(\mathit{Sodium}, \mathit{low})\}
\end{aligned}$$

Thus, test results denoted by the predicate symbol '$\mathit{Symptom}$' are assumed to be absent. Note that when applying this version of the consistency definition, obtained by the definition of O^c, the defect $\mathit{Disorder}(\mathit{influenza})$ cannot be part of any diagnosis, because this would clash with the consistency condition. Although, on first thought, the set

$$\{\mathit{Disorder}(\mathit{pulmonary\_embolism})\}$$

may seem to represent a diagnosis, it turns out that there exists no diagnosis at al. The reason is that

$$\mathrm{CM} \cup \{\mathit{Disorder}(\mathit{pulmonary\_embolism})\} \nvDash \mathit{Sign}(\mathit{fever})$$

i.e., the covering condition fails to hold.

A second, alternative version of the theory is presented in [Console & Torasso, 1990b] and [Console & Torasso, 1991]. In these articles, the consistency condition is reformulated, by adopting another definition for the set O^c, as follows. The set $O^c \subseteq \mathrm{OBS}_N$ is defined by:

$$O^c = \{\neg\pi(t) \in \mathrm{OBS}_N \mid \pi(s) \in O, t \neq s\}$$

where π stands for a predicate symbol, and t and s are constants. The consistency condition remains the same, but its effects on the computation of a diagnosis differs, because of the altered definition of O^c. For the example diagnostic problem, the set O^c is equal to

$$O^c = \{\neg \mathit{Sign}(\mathit{hypertension}), \neg \mathit{Blood\_chemistry}(\mathit{Sodium}, \mathit{low})\}$$

Note that the literals $\neg\mathit{Symptom}(\mathit{cough})$ and $\neg\mathit{Symptom}(\mathit{headache})$ are missing from this set, because none of the literals in the set of observed findings O has '*Symptom*' as predicate symbol. Thus, the test results with respect to test '*Symptom*' are assumed to be unknown. A diagnosis in this case is $H = \{\mathit{Disorder}(\mathit{influenza}), \mathit{Disorder}(\mathit{pulmonary\_embolism})\}$, because

$$\mathrm{CM} \cup H \vDash \{\mathit{Sign}(\mathit{fever}), \mathit{Blood\_chemistry}(O_2\text{-}\mathit{level}, \mathit{low})\}$$

(in fact, the literal $\mathit{Symptom}(\mathit{cough})$ is also entailed), and O^c is consistent with CM and H. Note that $H = \{\mathit{Disorder}(\mathit{influenza}), \mathit{Disorder}(\mathit{pulmonary\_embolism})\}$ yields an inconsistency if taken as a hypothesis using the first version of the consistency condition. ◇

The intuitive basis of the two versions of the consistency condition in abductive diagnosis, yielded by different logical interpretations of findings not observed, can be clarified in terms of diagnostic problem solving as follows.[3] In the first version of the consistency condition, it is assumed that all findings associated with a defect, present in the real world, will be observed. If a finding is not included among the findings in the set of observed findings, it is assumed to be absent; absent findings are denoted by negative literals. The basic assumption is that all findings of defects that are absent will not be observed, i.e. are absent (if unique for the defect), hence, it can safely be assumed that all findings not observed are negative. Although this may not be justified in diagnostic problem solving – it could be more natural to take the findings as unknown – the assumption of the negative literals has the technical advantage of blocking the inclusion of defects that are not present in the real world according to the theory, because some observable finding associated with the defect is not included in the set of observed findings. This is precisely the effect required. Now, if, as in the example above, only part of the unique findings of a defect occurs among the set of observed findings, there must be something wrong, either with the abnormality axioms CM, or with the set of observed findings. It seems therefore justified that no diagnosis is established in this case. However, this result is only valid if one accepts as a basic assumption that every possible cause (defect) of a finding is included in the set of abnormality axioms CM, which also constituted the basis of the predicate completion discussed above (at the risk of ambiguity with respect to database theory, one might call this the closed world assumption of abduction).

The second version of the consistency condition in abductive diagnosis is similar to the first version, except that it is assumed that if no information concerning a specific diagnostic test is available,– recall that every test corresponds to a different predicate symbol – it is assumed to be unknown. Now, if some defect d is included in a solution H and

$$\mathrm{CM} \cup \{d\} \vDash o$$

[3] We remark that this interpretation is the author's own, no such interpretation appears in the papers by Console and Torasso.

where $o \notin O$, this means that the model predicts that if the test is actually carried out, the finding o will be observed. If it is not observed, or turns out to be false, i.e. $\neg o$, some action needs to be undertaken, but no specific ideas concerning this situation appear in the papers of Console and Torasso. However, if the test has been carried out, i.e. there exists some finding o' with the same predicate symbol as o, and $o \notin O$, then again no diagnosis exists, because $\neg o \in O^c$ would hold.

The abductive theory of diagnosis discussed above may be viewed as a formalisation of particular parts of the expert system shell CHECK [Console & Torasso, 1989; Torasso & Console, 1989]. This system can be used to build hybrid diagnostic systems for domains in which causal, hierarchical and heuristic knowledge coexist. As far as known to the author, CHECK has been used as an experimental platform on which various prototype systems have been developed, including diagnosis of automobile engine failure and diagnosis of liver disease.

4.3 Set-covering theory of diagnosis

Instead of choosing logic as the language for MAB diagnosis, as discussed above, others have adopted set theory as their formal language. This approach to the formalisation of diagnosis is referred to as the *set-covering theory of diagnosis*, or *parsimonious covering theory* [Reggia et al., 1983; Allemang et al., 1987; Peng & Reggia, 1990; Wu, 1991]. The treatment of the set-covering theory of diagnosis in the literature deals only with the modelling of restricted forms of abnormal behaviour of a system.

The specification of the knowledge involved in diagnostic problem solving consists of the enumeration of all findings that may be present (and observed) given the presence of each individual defect distinguished in the domain; the association between each defect and its associated set of observable findings is interpreted as an uncertain *causal relation* between the defect and each of the findings in the set of observable findings. Instead of the terms 'defect' and 'finding' the terms 'disorder' and 'manifestation' are employed in descriptions of the set-covering theory of diagnosis. In the following, we have chosen to uniformly employ the terms 'defect' and 'finding' instead. The basic idea of the theory with respect to diagnosis is that each finding in the set of observed findings in a given diagnostic situation must be causally related to at least one present defect; the collected set of present defects thus obtained can be taken as a diagnosis. As with the theory of diagnosis by Console and Torasso, this reasoning method is usually viewed as being abductive in nature, because the reasoning goes from findings to defects, using causal knowledge from defects to findings.

More formally, the triple $\mathcal{N} = (\mathrm{DFS}, \mathrm{OBS}, C)$ is called a *causal net* in the set-covering theory of diagnosis, where

- DFS is a set of possible *defects*,
- OBS is a set of elements called *observable findings*, and
- C is a binary relation

  $$C \subseteq \mathrm{DFS} \times \mathrm{OBS}$$

  called the *causation relation*.

A *diagnostic problem* in the set-covering theory of diagnosis is then defined as a pair $\mathcal{D} = (\mathcal{N}, O)$, where $O \subseteq \mathrm{OBS}$ is a *set of observed findings*. It is assumed that all defects $d \in \mathrm{DFS}$

are potentially present in a diagnostic problem, and all findings $o \in \mathrm{OBS}$ will be observed when present. In addition, all defects $d \in \mathrm{DFS}$ have a causally related observable findings $o \in \mathrm{OBS}$, and vice versa, i.e. $\forall d \in \mathrm{DFS}\ \exists o \in \mathrm{OBS} : (d, o) \in C$, and $\forall o \in \mathrm{OBS}\ \exists d \in \mathrm{DFS} : (d, o) \in C$. No explicit distinction is made in the theory between positive (present), negative (absent) and unknown defects, and positive (present), negative (absent) and unknown findings. The causation relation is often depicted by means of a labelled, directed acyclic graph, which, as $\mathcal{N}$, is called a *causal net* [Peng & Reggia, 1990].

Let $\wp(X)$ denote the power set of the set X. It is convenient to write the binary causation relation C as two functions. Since in the next section, such functions are intensively employed, we adopt a notation that slightly generalises the notation proposed in [Peng & Reggia, 1990]. The first function

$$e : \wp(\mathrm{DFS}) \to \wp(\mathrm{OBS})$$

called the *effects function*, is defined as follows; for each $D \subseteq \mathrm{DFS}$:

$$e(D) = \bigcup_{d \in D} e(\{d\}) \tag{11}$$

where

$$e(\{d\}) = \{o \mid (d, o) \in C\}$$

and the second function

$$c : \wp(\mathrm{OBS}) \to \wp(\mathrm{DFS})$$

called the *causes function*, is defined as follows; for each $O \subseteq \mathrm{OBS}$:

$$c(O) = \bigcup_{o \in O} c(\{o\})$$

where

$$c(\{o\}) = \{d \mid (d, o) \in C\}$$

Hence, knowledge concerning combinations of findings and defects is taken as being composed of knowledge concerning individual defects or findings, which is not acceptable in general. This is a strong assumption, because it assumes that no interaction occurs between defects. In [Peng & Reggia, 1990], no attention is given to this subject.

A causal net can now be redefined, in terms of the effects function e above, as a triple $\mathcal{N} = (\mathrm{DFS}, \mathrm{OBS}, e)$.

Given a set of observed findings, diagnostic problem solving amounts to determining sets of defects – technically the term *cover* is employed – that account for *all* observed findings. Formally, a diagnosis is defined as follows.

Definition 6 (*set-covering diagnosis*). *Let $\mathcal{D} = (\mathcal{N}, O)$ be a diagnostic problem, where $\mathcal{N} = (\mathrm{DFS}, \mathrm{OBS}, e)$ is a causal net and O denotes a set of observed findings. Then, a* (set-covering) diagnosis *of $\mathcal{D}$ is a set of defects $D \subseteq \mathrm{DFS}$, such that:*

$$e(D) \supseteq O \tag{12}$$

In the set-covering theory of diagnosis the technical term 'cover' is employed instead of 'diagnosis'; 'diagnosis' will be the name adopted in this article. Due to the similarity of condition (12) with the covering condition in the abductive theory of diagnosis, this condition is called the *covering condition* in the set-covering theory of diagnosis. Actually, set-covering diagnosis can be mapped to abductive diagnosis in a straightforward way, thus revealing that set-covering diagnosis is more restrictive than abductive diagnosis. Just by mapping each function value

$$e(\{d\}) = \{o_1, \ldots, o_n\}$$

to a collection of logical implications, taken as abnormality axioms CM of a causal specification $\mathcal{C} = (\mathrm{DFS}, \mathrm{OBS}, \mathrm{CM})$, of the following form:

$$\begin{array}{rcl} d \wedge \alpha_{o_1} & \rightarrow & o_1 \\ d \wedge \alpha_{o_2} & \rightarrow & o_2 \\ & \vdots & \\ d \wedge \alpha_{o_n} & \rightarrow & o_n \end{array}$$

abductive diagnosis for such restricted causal specifications and set-covering diagnosis coincide.

Since it is assumed that $e(\mathrm{DFS}) = \mathrm{OBS}$ is satisfied, i.e. any finding $o \in \mathrm{OBS}$ is a possible causal effect of at least one defect $d \in \mathrm{DFS}$, there exists a diagnosis for any set of observed findings O, because

$$e(\mathrm{DFS}) \supseteq O$$

always holds (explanation existence theorem, [Peng & Reggia, 1990]).

A set of defects D is said to be an *explanation* of a diagnostic problem $\mathcal{D} = (\mathcal{N}, O)$, with O a set of observed findings, if D is a diagnosis of O and D satisfies some additional criteria. Various criteria, in particular so-called *criteria of parsimony*, are in use. The basic idea is that among the various diagnoses of a set of observable findings, those that satisfy certain criteria of parsimony are more likely than others. Let $\mathcal{D} = (\mathcal{N}, O)$ be a diagnostic problem, then some of the criteria as mentioned in [Peng & Reggia, 1990; Tuhrim et al., 1991] are:

- *Minimal cardinality*: a diagnosis D of O is an explanation of $\mathcal{D}$ iff it contains the minimum number of elements among all diagnoses of O;
- *Irredundancy*: a diagnosis D of O is an explanation of $\mathcal{D}$ iff no proper subset of D is a diagnosis of O;
- *Relevance*: a diagnosis D of O is an explanation of $\mathcal{D}$ iff $D \subseteq c(O)$;
- *Most probable diagnosis*: a diagnosis D of O is an explanation of $\mathcal{D}$ iff $P(D|O) \geq P(D'|O)$ for any diagnosis D' of O.

In addition, in [Charniak & Shimony, 1994], [Santos, 1994] and [Santos & Santos, 1996] minimal-cost diagnosis is defined. A diagnosis D of a set of observed findings O is called a *minimal-cost explanation* of $\mathcal{D}$ iff

$$\sum_{d \in D} cost(d) \leq \sum_{d \in D'} cost(d)$$

for each diagnosis D' of O, where *cost* is a function associating real values with defects $d \in \mathrm{DFS}$. The cost of a diagnosis may be anything, varying from financial costs to some subjective feeling of importance expressed by numbers. However, [Charniak & Shimony, 1994] choose as a semantics of cost function information for the negative logarithm of probabilities. Under this interpretation, a minimal-cost diagnosis is identical to a most probable diagnosis.

Although not every diagnosis is an explanation, any diagnosis may be seen as a solution to a diagnostic problem, where diagnoses which represent explanations conform to more strict conditions than diagnoses that do not. The term 'explanation' refers to the fact that a diagnosis in the set-covering theory of diagnosis can be stated, and thus be explained, in terms of cause-effect relationships. A better choice, in our opinion, would have been the adoption of the term 'explanation' for what is now called 'cover' in the theory, and to refer to what are now called 'explanations' by the name of 'parsimonious explanations'. To avoid confusion, the term 'explanation' will not be used in the sequel. Instead, we shall speak of a 'minimal-cardinality diagnosis', an 'irredundant diagnosis', a 'minimal-cost diagnosis' and so on.

For minimal cardinality, a diagnosis which consists of the smallest number of defects among all diagnoses is considered the most plausible diagnosis. Minimal cardinality is a suitable parsimony criterion in domains in which large combinations of defects are unlikely to occur. For example, in medicine, it is generally more likely that a patient has a single disorder than more than one disorder. Irredundancy expresses that it is not possible to leave out a defect from an explanation without losing the capability of explaining the complete set of observed findings, i.e.

$$e(D) \not\supseteq O$$

for each $D \subset D'$, where D' is an irredundant diagnosis. The relevance criterion states that every defect in an explanation has at least one observable finding in common with the set of observed findings. This seems an obvious criterion, but note that the notion of uncertain causal relation employed in the set-covering theory of diagnosis does not preclude situations in which a defect is present, although none of its causally related observable findings have been observed. These three definitions of the notion of explanation are based on general set-theoretical considerations. In contrast, the most probable diagnosis embodies some knowledge of the domain, in particular with respect to the strengths of the causal relationships. We shall not deal with such probabilistic extensions of the set-covering theory of diagnosis any further.

Example 7. Consider the causal net $\mathcal{N} = (\mathrm{DFS}, \mathrm{OBS}, C)$, where the effects function e is defined by the causation relation C, i.e.

$$e(D) = \bigcup_{d \in D} e(\{d\})$$

where

$$e(\{d\}) = \begin{cases} \{cough, fever, sneezing\} & \text{if } d = influenza \\ \{cough, sneezing\} & \text{if } d = common\ cold \\ \{fever, dyspnoea\} & \text{if } d = pneumonia \end{cases}$$

It states, for example, that a patient with influenza will be coughing, sneezing and have a fever; a patient with a common cold will show the same findings, except fever, and a patient with pneumonia will have a fever and dyspnoea (shortness of breath). The associated graph

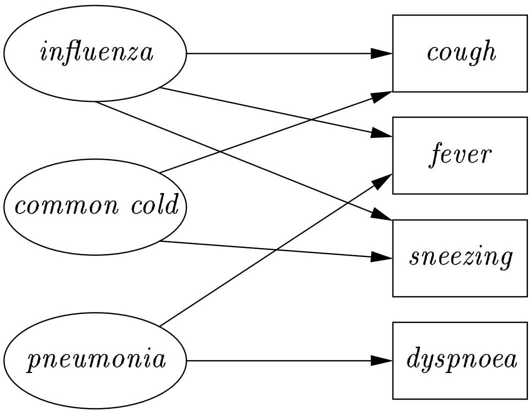

Figure 7: Causal net.

representation G_C of C is shown in Figure 7. It holds, among others, that

$$e(\{\mathit{influenza}, \mathit{common\ cold}\}) = \{\mathit{cough}, \mathit{fever}, \mathit{sneezing}\}$$

Based on the causal net C, the following causes function c is obtained:

$$c(O) = \bigcup_{o \in O} c(\{o\})$$

with

$$c(\{o\}) = \begin{cases} \{\mathit{influenza}, \mathit{common\ cold}\} & \text{if } o = \mathit{cough} \\ \{\mathit{influenza}, \mathit{pneumonia}\} & \text{if } o = \mathit{fever} \\ \{\mathit{influenza}, \mathit{common\ cold}\} & \text{if } o = \mathit{sneezing} \\ \{\mathit{pneumonia}\} & \text{if } o = \mathit{dyspnoea} \end{cases}$$

Suppose $\mathcal{D} = (\mathcal{N}, O)$ is a diagnostic problem, with $O = \{\mathit{cough}, \mathit{fever}\}$ a set of observed findings, then a diagnosis of $\mathcal{D}$ is

$$D_1 = \{\mathit{influenza}\}$$

but

$$\begin{aligned} D_2 &= \{\mathit{influenza}, \mathit{common\ cold}\} \\ D_3 &= \{\mathit{common\ cold}, \mathit{pneumonia}\} \end{aligned}$$

and $D_4 = \{\mathit{influenza}, \mathit{common\ cold}, \mathit{pneumonia}\}$ are also diagnoses for O. All of these diagnoses are relevant diagnoses, because

$$c(\{\mathit{cough}, \mathit{fever}\}) \supseteq D_i$$

where $i = 1, \ldots, 4$. Irredundant diagnoses of O are D_1 and D_3. There is only one minimal cardinality diagnosis, viz. $D_1 = \{\mathit{influenza}\}$. Now suppose that $O = \{\mathit{cough}\}$, then for example $D = \{\mathit{influenza}, \mathit{pneumonia}\}$ would not have been a relevant diagnosis, because

$$c(\{\mathit{cough}\}) = \{\mathit{influenza}, \mathit{common\ cold}\} \not\supseteq D$$

$\diamond$

Other, more domain-specific, definitions of the notion of explanation have only been developed recently. Such domain-specific knowledge can be effective in reducing the size of the set of diagnoses generated by a diagnostic system. For example, [Tuhrim et al., 1991] demonstrated that the use of knowledge concerning the three-dimensional structure of the brain by means of a binary adjacency relation in a neurological diagnostic expert system, based on the set-covering theory of diagnosis, could increase the diagnostic accuracy of the system considerably.

In [Peng & Reggia, 1990], it is shown that the causation relation C can be extended for the representation of multi-layered causal nets, in which defects are causally connected to each other, finally leading to observable findings. By computation of the reflexive, transitive closure of the causation relation, $C^{\star}$, the basic techniques discussed above immediately apply. The reflexive closure makes it possible to enter defects as observed findings, which are interpreted as already established defects, yielding a slight extension to the theory treated above.

In the set-theoretical formalisation of diagnosis by Bylander et al., [Bylander et al., 1992], an effects function e is used to represent both the knowledge base and the method of diagnostic problem solving. In contrast to the theory by Peng and Reggia, the function e can be used to represent diagnostic interactions among defects, because the assumption that $e(D)$ is the union of function values $e(\{d\})$, for each $d \in D$, is not generally assumed. A diagnosis D is defined by $e(D) = O$, where O is a set of observed findings, i.e. every finding must be covered by the set of defects D, which is very restrictive.

Charniak and Shimony, [Charniak & Shimony, 1994], and Santos, [Santos, 1994], generalise set-covering theory by representing causal knowledge as directed AND/OR graphs. Hence, various types of causal interaction can be represented. A diagnosis is defined as a minimal-cost solution, i.e. they restrict to cost-based abduction.

INTERNIST-1/QMR is an example of an expert system with a basis related to the set-covering theory [Miller et al., 1982; Bankowitz et al., 1989; Peng & Reggia, 1990]. The system is not a direct implementation of the theory reviewed above; in fact, the system predates the theory for about a decade. However, knowledge in INTERNIST-1/QMR is organised in a way very similar to that of set-covering theory, and a diagnosis produced by the system bears great resemblance to a diagnosis in set-covering theory. It deviates from this theory in several respects, in particular by employing domain-specific heuristics in selecting diagnoses [Peng & Reggia, 1990]. RED is an expert system in the domain of blood bank antibody analysis [Josephson & Josephson, 1994; Punch III et al., 1990; Smith et al., 1985]. This system can also be described in terms of the set-covering theory of diagnosis, although several aspects of the system go beyond the theory, such as the representation of interactions among particular antibody reactions, requiring a generalisation of the set-covering theory [Bylander et al., 1992] (See also Section 5). Peirce is a domain-independent tool that generalised on the techniques used in RED [Punch III et al., 1990]. In [Tuhrim et al., 1991], an expert system for the diagnosis of brain lesions, based on the set-covering theory of diagnosis is described. Of the systems mentioned above, the last system is based most clearly on the principles described in this section.

4.4 Hypothetico-deductive diagnosis

The third approach to diagnosis mentioned in Section 3, AC (Abnormality Classification) diagnosis, originates from work by E.H. Shortliffe, B.G. Buchanan, W.J. Clancey and E.A. Feigenbaum in the MYCIN project [Shortliffe, 1976; Buchanan & Shortliffe, 1984; Clancey & Letsinger, 1984]. The knowledge incorporated in that expert system, and in similar systems for AC diagnosis, is based on the body of experience accumulated in handling a large number of cases, such as the patients a physician sees in medical practice. The knowledge is extracted from textbooks or human experts. We have called this type of knowledge empirical associations, i.e. the knowledge consists of associations between typical observable findings and defects; knowledge about the underlying mechanisms (if available) is not represented.

In most practical systems (e.g. the HEPAR system, [Lucas & Janssens, 1991]), the formal counterparts of empirical associations are organised according to some underlying model distinguished in the collection of empirical associations. A typical example is a distinction between families of disorders and specific disorders, i.e. a taxonomy of disorders, that can be exploited in problem solving. Hence, expert systems based on empirical associations are model-based like the other systems discussed above, because they are also based on a model of the problem domain, although the nature of the model is different. It is possible to characterise AC diagnosis in a more formal way. We shall refer to this formal counterpart of AC diagnosis as *hypothetico-deductive diagnosis*, a term suggested in [Campbell, 1976] and [Macartney, 1988].

A hypothetico-deductive diagnostic problem consists of a set of logical axioms, called an *empirical model* EM, of the form

$$c_1 \wedge \cdots \wedge c_n \to q \tag{13}$$

where c_i and q represent either negative or positive defects and findings, represented in logic as negative or positive literals, and if every c_i is a finding, then q should be a defect. Logical implication in the formalisation of empirical associations (13) may be viewed as a *classification relation*. A set of observed findings is represented as a set of ground literals, where each literal is of the finding type. For example, a typical logical axiom might be

$$o_1 \wedge \cdots \wedge o_m \to d$$

which expresses that a set of observable findings $O = \{o_1, \ldots, o_m\}$ represents necessary and sufficient evidence for establishing the presence of the defect d as part of a diagnosis. One difference between the theories of hypothetico-deductive diagnosis and abductive diagnosis is that, in hypothetico-deductive diagnosis, observed findings and defects need not be causally related to each other. Some of the findings may be interpreted as abnormal; other findings, such as, for example, age of a patient in a medical application, may not. The function of normal findings in empirical associations is similar to that of conditional causality introduced in Section 4.2, viz. to condition a particular piece of knowledge on a specific piece of evidence.

Now, let $\mathcal{B} = (\mathrm{DFS}, \mathrm{OBS}, \mathrm{EM})$ denote an *associational specification*, where:

- DFS denotes a set of (positive and negative) possible defects,
- OBS denotes a set of (positive and negative) observable findings, and
- EM denotes the logical representation of a set of empirical associations of the form (13).

A *hypothetical-deductive diagnostic problem* is then defined as a pair $\mathcal{H} = (\mathcal{B}, O)$, where $O \subseteq \mathrm{OBS}$ denotes a *set of observed findings*. A diagnosis based on empirical associations can be defined as follows.

Definition 7 (*hypothetico-deductive diagnosis*). *Let $\mathcal{H} = (\mathcal{B}, O)$ be a hypothetico-deductive diagnostic problem, where $\mathcal{B} = (\mathrm{DFS}, \mathrm{OBS}, \mathrm{EM})$ is an associational specification, and O is a*

Originator	Knowledge base specification	Knowledge base interpretation	Diagnosis
Reiter	functional relations	deduction	consistency
Console & Torasso	causality	abduction deduction	covering consistency
Reggia et al.	causality	abduction	set covering
Bylander et al.	diagnostic relation	none	set covering
Shortliffe et al.	empirical associations	deduction	classification

Table 2: Comparison of formal theories of diagnosis.

set of observed findings. Let $\Theta \subseteq \mathrm{DFS}$ *be a set of defects, called a* hypothesis. *Then,* $D \subseteq \Theta$ *is called a* (hypothetico-deductive) diagnosis *of* $\mathcal{H}$ *if*

$$D = \{d \in \Theta \mid \mathrm{EM} \cup O \vDash d\}$$

Note that, in contrast with the theories discussed above, a single hypothesis is initially given in hypothetico-deductive diagnosis; it stands for the defects that are initially given to be of interest. In the theory of hypothetico-deductive diagnosis, defects are logically entailed by the observed findings (usually implemented by a deductive calculus, hence the adjective hypothetico-*deductive*).

In contrast with the other theories of diagnosis, there are a large number of nonexperimental applications available that may be viewed as hypothetico-deductive diagnostic systems.

The technical characteristics of the various formal theories of diagnosis, discussed in the previous sections, are summarised in Table 2.

5 Frameworks of diagnosis

Having described the various formal theories of diagnosis, the question arises in what sense these theories are related to each other, and whether it is possible to develop generalisations based on these theories. Actually, several originators of theories of diagnosis have investigated the expressiveness of their theory for modelling other conceptual models of diagnosis than those for which the theory was originally designed. In this section, we summarise and comment on results found in the literature, and discuss various general frameworks of diagnosis.

5.1 Expressiveness of theories of diagnosis

Reiter has shown that the framework of consistency-based diagnosis provides enough descriptive power to capture the set-covering theory of diagnosis [Reiter, 1987]. In Reiter's formalisation, the normality axioms in the original theory of consistency-based diagnosis are changed into *abnormality axioms*, simply by replacing 'components' by 'defects'. These axioms have the following form

$$\neg Abnormal(d) \rightarrow \neg Present(d) \tag{14}$$

for each defect d, stating that under normal conditions defect d is not present, and

$$o_{ab} \rightarrow Present(d_1) \vee \cdots \vee Present(d_n) \tag{15}$$

for each observable abnormal finding o_{ab} and related defect d_i, $i = 1, \ldots, n$. Formulae of the form (14) express hypotheses, namely that a particular defect may be absent ($\neg Present(d)$) if it does not give rise to an inconsistency. As we have discussed in Section 4.2, formulae of the form (15) may be seen as the predicate completion [Clark, 1978], of finding literals in formulae of the form

$$Present(d) \to o_{ab}$$

i.e. if CM denotes the set of formulae of the last form, with defect literals $Present(d_1)$, ..., $Present(d_n)$ in the premise, then the predicate completion $\mathrm{COMP}[\mathrm{CM}; o_{ab}]$ with regard to the finding o_{ab} is equal to

$$\mathrm{COMP}[\mathrm{CM}; o_{ab}] = \mathrm{CM} \cup \{o_{ab} \to Present(d_1) \vee \cdots \vee Present(d_n)\}$$

This states that the only causes of the finding o_{ab} to be present (and observed) are the defects $d_1, \ldots, d_n$. As discussed above, this same kind of knowledge is expressed, although implicitly, in the abductive theory of diagnosis; it is also expressed in the set-covering theory of diagnosis, but the differences between the reasoning methods employed (consistency-based reasoning, logical abduction, and set covering) dictate a different representation (syntax) in all three formal theories. Informally, in the consistency-based diagnosis formalisation of MAB diagnosis, diagnostic problem solving is carried out as follows. Given an observed finding o_{ab} associated with a defect d_i, $i = 1, \ldots, n$, a disjunction

$$Present(d_1) \vee \cdots \vee Present(d_n)$$

is deduced, which is reduced by cancelling out atoms using axiom (14), assuming certain defects not to be present, i.e. *Abnormal*(d) is *false*, yielding a (subset minimal) diagnosis. The effect of axiom (14) corresponds to producing irredundant diagnoses in the set-covering theory of diagnosis, in the sense that a minimal diagnosis with respect to set inclusion is produced. Reiter shows that there exists a (subset minimal) diagnosis according to the consistency-based reformulation of the set-covering theory of diagnosis iff there exists an equivalent irredundant diagnosis in the set-covering theory (although at the time Reiter's result was published, the notion of irredundant diagnosis had not yet appeared in the literature) [Reiter, 1987].

Console and Torasso have studied the use of the consistency condition in abductive diagnosis for modelling DNSB diagnosis, i.e. diagnosis using a specification of a model of normal structure and behaviour in a way resembling the work of Reiter [Reiter, 1987; Console & Torasso, 1990b; Console & Torasso, 1991]. By taking the empty set for the set of observed findings that must be covered, the covering condition in abductive diagnosis becomes

$$\mathrm{CM} \cup H \models O'$$

where $O' = \varnothing$; a diagnosis is the result of satisfaction of the consistency condition only, because the covering condition is always satisfied in this case. Thus, consistency-based diagnosis in the sense of Reiter is obtained. However, the meaning of the logical axioms is entirely different from the meaning originally attached to the logical axioms, because they now represent normal behaviour of a device; d represents some normal state of a component of the device and a finding o in the conclusion of a Horn clause $d \to o$ represents a finding that may be observed when the component is in its normal state, i.e. o represents a normality finding o_{norm}. By varying between $O' = \varnothing$ and $O' = O$, for example by taking for O' the set of all abnormal findings o_{ab} occurring in O, DNSB and MAB diagnosis can be integrated within the same abductive framework [Console & Torasso, 1990b; Console & Torasso, 1991]. The resulting abductive framework is referred to as 'the spectrum of logical definitions of diagnosis'.

We may conclude by saying that generalisation of the formal theories of diagnosis discussed above has shown that there is no such thing as a unique formalisation of a conceptual model of diagnosis. Although the formal theories can be applied to formalise conceptual models of diagnosis other than those for which they were originally designed, the results often lack conceptual clarity.

5.2 Generalisation towards frameworks of diagnosis

The principal difficulty of developing a theory of diagnosis lies, undoubtedly, in the design of a mapping of some intuitively appealing conceptual model of diagnosis to a formal language, such as logic or set theory. We know beforehand that both logic and set theory are sufficiently expressive [Lewis & Papadimitriou, 1981]; so, this is not where the problem lies. The selection of an appropriate logic, or an appropriate fragment of set theory, however, is much more difficult. The insights gained from the formal theories discussed in Section 4 have facilitated researchers in coming up with more general frameworks of diagnosis.

In [Ten Teije & Van Harmelen, 1994] it is proposed to extend the spectrum of logical definitions of diagnosis, discussed in Section 5.1, by leaving the choice of the relations for defining the covering and consistency conditions open (one trivial possibility would be to choose the logical entailment relation $\vDash$ as a basis for both relations, another choice would be an approximate entailment relation, such as defined by Schaerf and Cadoli [Schaerf & Cadoli, 1995; Ten Teije & Van Harmelen, 1996]), and by making it possible to choose an arbitrary decomposition of the set of observed findings O into O^c, the set that must be consistent with a diagnosis, and O', the set that must be covered by a diagnosis. As a consequence, the resulting framework is more flexible than the original framework, although it is still in the spirit of the original spectrum of logical definitions of diagnosis. While the framework is tailored to diagnosis, devising suitable definitions for the covering and consistency relations is far from trivial.

D. Poole and colleagues have developed a theory and an implementation of a form of hypothetical reasoning, called *Theorist* [Poole et al., 1987]. Theorist may be used as a framework of diagnosis, but it is not restricted in any way to diagnostic problem solving. Moreover, there are no inherent relationships between Theorist and any of the conceptual models of diagnosis. The present implementation of the Theorist framework, however, is more or less tailored to abductive diagnosis.

In Theorist, a diagnostic problem must be specified in terms of a set of *facts*, denoted by FACTS, a set of *hypotheses*, denoted by HYP, and a set of *constraints*, denoted by C. The set of facts FACTS and constraints C are collections of arbitrary closed formulae in first-order logic; hypotheses act as a kind of defaults that might become instantiated, and assumed to hold true, in the reasoning process. A set FACTS $\cup H$ is called an *explanation* of a closed formula g, where H is a set of ground instances of hypothesis elements in HYP, iff:

(1) $\mathrm{FACTS} \cup H \vDash g$, and

(2) $\mathrm{FACTS} \cup H \cup C \nvDash \bot$.

On first sight, the framework looks a lot like the framework of abductive diagnosis discussed in Section 4.2, but it is much more general, mainly due to the unrestricted nature of its

```
default a1.
default a2.
default fever.
default influenza.
default sport.

fact chills  <- fever and a1.
fact fever   <- influenza.
fact thirst  <- fever.
fact myalgia <- influenza and a2.
fact myalgia <- sport.

constraint not chills.  % Oc
```

Figure 8: Specification of an abductive diagnostic problem in Theorist.

elements. In terms of the abductive theory of diagnosis, we would have called H a solution, if the abnormality axioms CM were taken as FACTS, the set of findings not observed O^c as constraints C, and the set of observed findings O as g. Obviously, because there is no fixed diagnostic interpretation in Theorist, the framework can be used as a basis for various other notions of diagnosis, such as consistency-based diagnosis (just take $g \equiv \top$). A similar framework of diagnosis has been proposed by K. Konolige [Konolige, 1994]; in this theory, called the *default causal net theory*, the partitioning of a logical theory into various meaningful elements is a bit different from that of Theorist. However, there are too many similarities with Theorist to justify an extensive description in this overview.

Example 8. Figure 8 presents a specification of the abductive diagnostic problem from Example 3 in terms of the Theorist implementation [Poole, 1990c], where `Oc` denotes the set of findings assumed to be absent, O^c, taken as constraints in Theorist. The following query:

```
explain thirst and myalgia.
```

yields the following results:

```
Answer is thirst and myalgia
Theory is [a2,influenza,fever]

Answer is thirst and myalgia
Theory is [sport,fever]

Answer is thirst and myalgia
Theory is [a2,influenza]

Answer is thirst and myalgia
Theory is [sport,influenza]
```

Theories are solutions in the abductive theory of diagnosis. Only a subset of the solutions mentioned in Example 3 are computed, because in Theorist it is assumed that every observed

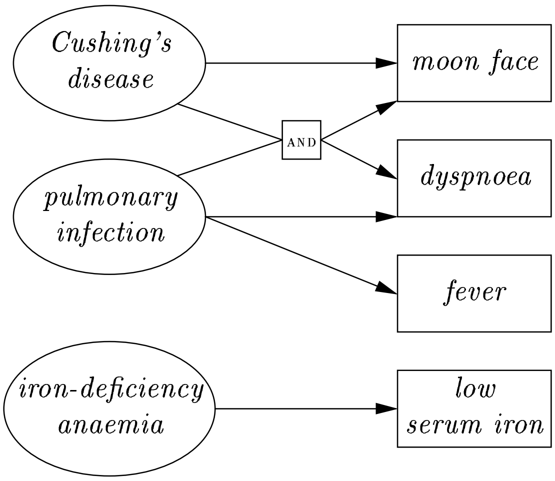

Figure 9: Nonmonotonic interaction between disorders.

finding need be explained only once by a diagnosis. ◇

Although logic offers powerful tools for designing and studying notions of diagnosis, formalising diagnosis in logical terms may be cumbersome. In particular, when it is necessary to resort to non-standard logics, there is a lurking danger that the original subject of research (diagnosis) is taken over by the study of logic. Instead of taking logic as a language to develop a framework of diagnosis, we might also adopt set theory as our language of choice and generalise the set-covering theory of diagnosis. This approach, which is more straightforward than logical analysis, has been investigated in [Lucas, 1996a]. The approach is introduced by the following example.

Example 9. Consider a medical diagnostic problem, where a patient may have Cushing's disease – a disease caused by a brain tumour producing hyperfunctioning of the adrenal glands – pulmonary infection and iron-deficiency anaemia. We shall not enumerate all symptoms and signs causally associated with these medical problems; it suffices to note that moon face is a sign associated with Cushing's disease, fever and dyspnoea (shortness of breath) are associated with pulmonary infection, and low levels of serum iron are characteristic for iron-deficiency anaemia. However, in a patient in whom Cushing's disease and pulmonary infection coexist there usually is no fever. This indicates that there exists an interaction between the two disorders, Cushing's disease and pulmonary infection, that is nonmonotonic, i.e. the co-occurrence of the two disorders produces fewer findings than the union of their associated observable findings. Figure 9 depicts this simple problem. Note that we can neither represent this knowledge by a causal specification (refraining from non-standard logic) as used in abductive diagnosis, nor in terms of an effects function as used in the set-covering theory of diagnosis. ◇

Interactions among defects (disorders) can be expressed by means of a mapping of sets of defects to sets of observable findings. Such a mapping will be called an evidence function. More formally, let $\Sigma = (\mathrm{DFS}, \mathrm{OBS}, e)$ be a *diagnostic specification*, where, again, DFS denotes a set of possible defects (disorders), and OBS denotes a set of observable findings. Positive defects d (findings o) and negative defects $\neg d$ (findings $\neg o$) denote *present* defects (findings) and *absent* defects (findings), respectively. If a defect d or a finding o is not included in a set, it is assumed to be *unknown*. Let a set X_P denote a set of positive elements, and let X_N denote a set of negative elements, such that X_P and X_N are disjoint. It is assumed that $\mathrm{DFS} = \mathrm{DFS}_P \cup \mathrm{DFS}_N$ and $\mathrm{OBS} = \mathrm{OBS}_P \cup \mathrm{OBS}_N$. Now, an *evidence function* e is a mapping

$$e : \wp(\mathrm{DFS}) \rightarrow \wp(\mathrm{OBS}) \cup \{\perp\}$$

such that:

(1) for each $o \in \mathrm{OBS}$ there exists a set $D \subseteq \mathrm{DFS}$ with $o \in e(D)$ or $\neg o \in e(D)$ (and possibly both);

(2) if $d, \neg d \in D$ then $e(D) = \perp$;

(3) if $e(D) \neq \perp$ and $D' \subseteq D$ then $e(D') \neq \perp$.

If $e(D) \neq \perp$, it is said that $e(D)$ is the set of *observable findings* for D; otherwise, it is said that D is inconsistent. Inconsistency here means that a particular combination of defects is not allowed. According to the definition above, we may have that both $o \in e(D)$ and $\neg o \in e(D)$, which simply means that these findings may alternatively, e.g. at different times, occur given the combined occurrence of the defects in the set D.

For the medical knowledge depicted in Figure 9, it holds, among others, that:

$$\begin{aligned} e(\{\mathit{Cushing's\ disease}\}) &= \{\mathit{moon\ face}\} \\ e(\{\mathit{pulmonary\ infection}\}) &= \{\mathit{fever}, \mathit{dyspnoea}\} \\ e(\{\mathit{Cushing's\ disease}, \mathit{pulmonary\ infection}\}) &= \{\mathit{moon\ face}, \mathit{dyspnoea}\} \end{aligned}$$

The property

$$\begin{aligned} e(\{\mathit{Cushing's\ disease}, \mathit{pulmonary\ infection}\}) \not\supseteq\ & e(\{\mathit{Cushing's\ disease}\}) \cup \\ & e(\{\mathit{pulmonary\ infection}\}) \end{aligned}$$

formally expresses that the interaction between *Cushing's disease* and *pulmonary infection* is nonmonotonic.

Various semantic properties of a domain for which a diagnostic system must be built can be expressed precisely as interactions in terms of evidence functions. An example of a local interaction reflected in an evidence function is *causality*; it is formalised as $e(D') \subseteq e(D)$, with the following meaning: 'the set of defects D *causes* the set of defects D''.[4] This is the same sort of knowledge as used in abductive diagnosis (cf. Section 4.2). We may also have that defects exhibit no interactions at all, which is a *global* property, expressed as follows:

$$e(D) = \bigcup_{d \in D} e(\{d\})$$

for each consistent set $D \subseteq \mathrm{DFS}$. Observe that this evidence function corresponds to the effects function (11) in the set-covering theory of diagnosis. Other semantic properties (with respect to observable findings) can be defined in this fashion quite easily [Lucas, 1996a].

To employ an evidence function for the purpose of diagnosis, it must be interpreted with respect to the actually observed findings. The interpretation of an evidence function and the

[4] We do not claim that this property formalises causality; it only expresses the notion of causality in terms of diagnosis.

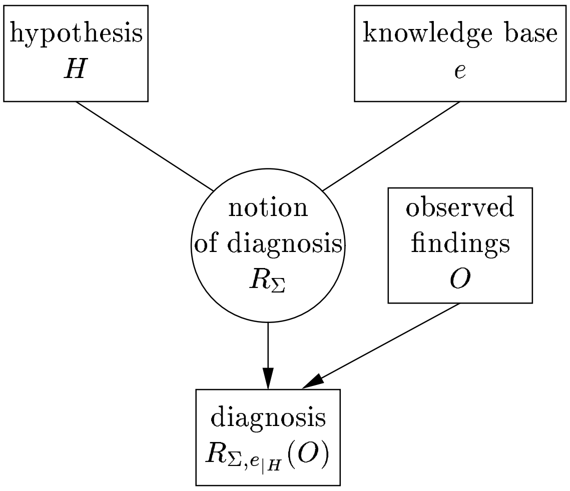

Figure 10: Schema of notion of diagnosis, diagnostic problem and solution.

observed findings that is adopted, can be viewed as a notion of diagnosis applied to solve the diagnostic problem at hand.

More formally, let $\mathcal{P} = (\Sigma, O)$ be a *diagnostic problem*, where $\Sigma = (\text{DFS}, \text{OBS}, e)$ and $O \subseteq \text{OBS}$ is a set of *observed findings*. Let R_Σ denote a *notion of diagnosis* R applied to Σ, then a mapping

$$R_{\Sigma,e_{|H}} : \wp(\text{OBS}) \to \wp(\text{DFS}) \cup \{u\}$$

will either provide a diagnostic solution for a diagnostic problem $\mathcal{P}$, or indicate that no solution exists, denoted by u (undefined). Here, H denotes a *hypothesis*, which is taken to be a set of defects ($H \subseteq \text{DFS}$), and $e_{|H}$, called the *restricted evidence function* of e, is a restriction of e with respect to the power set $\wp(H)$:

$$e_{|H} : \wp(H) \to \wp(\text{OBS}) \cup \{\bot\}$$

where for each $D \subseteq H$: $e_{|H}(D) = e(D)$. A restricted evidence function $e_{|H}$ can be thought of as the relevant part of a knowledge base with respect to a hypothesis H. An *R-diagnostic solution*, or *R-diagnosis* for short, with respect to a hypothesis $H \subseteq \text{DFS}$, is now defined as the set

$$R_{\Sigma,e_{|H}}(O), \text{ where } R_{\Sigma,e_{|H}}(O) \subseteq H \text{ if a solution exists.}$$

The general idea is illustrated in Figure 10. To illustrate the flexibility of the framework, consider again the notion of *weak* causality as defined in the abductive theory of diagnosis, which is obtained by the addition of assumption literals α to individual abnormality axioms of a causal model CM.

Example 10. Consider the abductive problem $\mathcal{A} = (\mathcal{C}, O)$, with causal specification $\mathcal{C} = (\text{DFS}, \text{OBS}, \text{CM})$, where CM is equal to:

$$\begin{array}{rcl}
\mathit{fever} \wedge \alpha_1 & \to & \mathit{thirst} \\
\mathit{fever} \wedge \alpha_2 & \to & \mathit{sweating} \\
\mathit{pneumonia} \wedge \alpha_3 & \to & \mathit{fever} \\
\mathit{pulmonary\_embolism} \wedge \alpha_4 & \to & \mathit{dyspnoea}
\end{array}$$

and where *fever*, *pneumonia*, and *pulmonary_embolism* are defects (disorders). The resulting evidence function e is defined by the following restriction $\tilde{e}$ of the evidence function e:

$$\tilde{e}(D) = \begin{cases} \{\mathit{thirst}, \mathit{sweating}\} & \text{if } D = \{\mathit{fever}\} \\ \{\mathit{thirst}, \mathit{sweating}\} & \text{if } D = \{\mathit{pneumonia}\} \\ \{\mathit{dyspnoea}\} & \text{if } D = \{\mathit{pulmonary\_embolism}\} \\ \bot & \text{if } D = \{\neg \mathit{fever}, \mathit{pneumonia}\} \\ \varnothing & \text{if } D \text{ is a singleton set different from those above} \end{cases}$$

yielding a diagnostic specification $\Sigma = (\mathrm{DFS}, \mathrm{OBS}, e)$, where the function e is obtained from $\tilde{e}$ by taking the union of non-specified, consistent function values. For example,

$$e(\{\mathit{pneumonia}, \mathit{pulmonary\_embolism}\}) = \{\mathit{thirst}, \mathit{sweating}, \mathit{dyspnoea}\}$$

◇

Given the definition of a diagnostic problem $\mathcal{P}$, it is possible to solve it using various notions of diagnosis. For example, the notion of diagnosis that corresponds to abductive diagnosis with weakly causal relations as introduced above, is called the notion of *weak-causality diagnosis*, denoted by WC. It is defined as follows:

$$\mathrm{WC}_{\Sigma, e_{|H}}(O) = \begin{cases} H & \text{if } e_{|H}(H) \supseteq O \\ u & \text{otherwise} \end{cases}$$

This notion of diagnosis is precisely the same as set-covering diagnosis, except that it is defined for general evidence functions, and not only for those evidence functions that are free of interaction.

Example 11. Reconsider the previous example. Let the set of observed findings be equal to $O = \{\mathit{thirst}, \mathit{sweating}\}$, then the set $H = \{\mathit{fever}, \alpha_1, \alpha_2\}$ is an abductive solution to $\mathcal{A} = (\mathcal{C}, O)$, because the covering and consistency conditions are satisfied; the associated diagnosis is $D = \{\mathit{fever}\}$. In terms of the set-theoretical framework, we have

$$\mathrm{WC}_{\Sigma, e_{|\{\mathit{fever}\}}}(O) = \{\mathit{fever}\}$$

Hence, the results of the (set-theoretical) notion of weak-causality diagnosis and the (logical) notion of abductive diagnosis with a weakly causal model CM do indeed coincide. ◇

Other notions of diagnosis, such as consistency-based diagnosis or a notion of diagnosis based on strongly causal knowledge, can be defined in a straightforward way. For example, where the notion of strong causality diagnosis is obtained in the theory of abductive diagnosis by doing away with incompleteness assumption literals, the same notion is obtained in the set-theoretical framework by replacing the $\supseteq$ relation in the definition of the function WC by equality $=$. The resulting notion of diagnosis expresses that all predicted observable findings must be observed, and vice versa.

It is also straightforward to define notions of diagnosis in terms of the set-theoretical framework that offer some approximating or refinement form of diagnosis. For example, the following notion of diagnosis, called *most general subset diagnosis* [Lucas, 1996b],

$$\mathrm{GS}_{\Sigma, e_{|H}}(E) = \begin{cases} \bigcup_{\substack{H' \subseteq H \\ e_{|H}(H') \subseteq E}} H' & \begin{array}{l} \text{if } H \text{ is consistent, and} \\ \quad \exists H' \subseteq H : e_{|H}(H') \subseteq E \end{array} \\ u & \text{otherwise} \end{cases}$$

is more flexible than strong-causality diagnosis. Intuitively, a most general subset diagnosis is the smallest set of defects that includes all accepted subhypotheses of a given hypothesis, where an accepted subhypothesis concerns observable findings that all have been observed.

Example 12. Reconsider Example 10. Let $O = \{\mathit{thirst}, \mathit{sweating}, \mathit{dyspnoea}\}$ be the set of observed findings. Then, we have that

$$\mathrm{WC}_{\Sigma, e_{|\{\mathit{fever},\mathit{pneumonia}\}}}(O) = u$$

i.e. the observed findings in O cannot be accounted for using weak-causality diagnosis. However, it holds that

$$\mathrm{GS}_{\Sigma, e_{|\{\mathit{fever},\mathit{pneumonia}\}}}(O) = \{\mathit{fever}, \mathit{pneumonia}\}$$

This expresses that at least part of the observed findings in O can be accounted for by the hypothesis $\{\mathit{fever}, \mathit{pneumonia}\}$. ◇

Hypothetico-deductive diagnosis can be described using the set-theoretical framework as a specific form of most general subset diagnosis. Assuming for simplicity's sake that the associated evidence function exhibits no interaction, most general subset diagnosis expresses that a defect is accepted as part of a diagnosis if all its associated typical observable findings have been observed. With some slight extensions, it is also possible to model the effect of grouping various findings with respect to a defect, which is usually expressed in rule-based systems by defining more than one rule with the same conclusion.

6 Discussion

The overview of the various approaches to diagnostic problem solving presented above indicates that, on the one hand, several different formalisations of the same conceptual model of diagnosis exist, whereas, on the other hand, several different conceptualisations of diagnosis fit into the same formal framework. Unfortunately, the various conceptual models of diagnosis presented in the literature are still commonly referred to by the name of their formal counterpart, suggesting that a unique linkage does exist between a formal theory and a conceptual model of diagnosis.

Each formal theory of diagnosis discussed has originally been developed to capture one specific conceptual approach to diagnosis. This remains visible, in spite of attempts of generalisation. They seem too intimately linked with their conceptual bases to be taken as genuine formal frameworks of diagnosis. The theory of consistency-based diagnosis as proposed by Reiter, [Reiter, 1987], and De Kleer et al., [De Kleer et al., 1992], provides a framework of both DNSB and MAB diagnosis, although the theory appears rather cumbersome for expressing MAB diagnosis. It does not provide a suitable basis for AC diagnosis. As the theory is based on the general, logical notion of (in)consistency, it is not expressive enough to capture many of the essential features of diagnostic problem solving in a straightforward way. Inconsistency may be a suitable notion to describe deviation from the normal situation of a device, but as a model for the description of the relationships among defects and findings it is highly unnatural.

In the abductive theory of diagnosis proposed by Console and Torasso, specific assumptions are made with respect to the causal nature of the knowledge involved. Their formalisation implicitly assumes that logical implication provides a suitable axiomatisation of the notion of causality. However, only the transitive nature of logical implication seems to meet the properties of causality; the reflexive and contrapositive properties of implication will not hold for all notions of causality. The interpretation of causal knowledge in the theory by Console and Torasso is achieved through the logical entailment relation, which is also used, together with the consistency and covering condition, to define notions of diagnosis. Hence, no clear distinction is made between the interpretation of a knowledge base – to determine what logically follows from the knowledge base – and applying this interpretation to determine a diagnosis. Furthermore, by the monotonicity of the entailment relation, certain types of knowledge, e.g. knowledge in which observable findings are cancelled due to interaction among defects, are precluded from formalisation. Extension to general Horn clause logic has been investigated by several researchers (cf. [Console et al., 1991] and [Preist et al., 1994]), but this is only one of the possibilities.

The set-covering theory of diagnosis aims, like the abductive theory of diagnosis, at describing a domain in terms of causality. Unlike the abductive theory of diagnosis, only a single concept of causality is employed, which is only made more expressive by the interpretation of causal relations as conditional probabilities [Peng & Reggia, 1990], yielding a formalism that is much alike the belief-network formalism [Lucas & Van der Gaag, 1991; Pearl, 1988]. Furthermore, a knowledge base consists only of a specification of single defects in terms of associated findings; in this way, it is not possible to model interactions among defects. Moreover, the notion of set-covering diagnosis (explanation) is fixed, with the exception of the notion of minimal diagnosis, which is variable in the theory. Finally, no clear distinction is made between the interpretation of a knowledge base in terms of causality and the process of diagnosis. We concluded that many of the restrictions underlying theories of diagnosis are too strong, certainly if they are to be used as frameworks of diagnosis.

The Theorist framework may be taken as a framework of diagnosis, but it seems more appropriate to view it as a general framework of hypothetical or default reasoning. Here, the relationship with diagnosis is actually too weak to accept it as a framework of diagnosis. The 'spectrum of logical definitions of diagnosis' by Console and Torasso, [Console & Torasso, 1991], with its generalisation by Ten Teije and Van Harmelen, [Ten Teije & Van Harmelen, 1994], is much more tailored to the encoding of conceptual models of diagnosis. Further extension of the spectrum to temporal model-based diagnosis has been proposed recently [Brusoni et al., 1996]. For defining hypothetico-deductive diagnosis, however, this approach is not very suitable.

Although logic has been adopted as a language for formalisation in most formal theories of diagnosis, set theory offers a powerful alternative. For example, the set-theoretical framework of diagnosis proposed in [Lucas, 1996a] leaves much freedom to the designer of a diagnostic theory, because it does not require the designer to comply with the constraints of some predefined semantics, such as underlying standard logic. Of course, it is usually desirable to define notions of diagnosis that closely mirror the meaning of a knowledge base. An advantage is that the framework supports the design of notions of diagnosis from scratch, and various kinds of interactions can be expressed readily in the framework. A disadvantage of the framework is that it is rather extensional in nature, hence less suitable for the modelling of domain knowledge.

Although during the last few years, research has focused on frameworks of diagnosis, it is not clear as yet whether frameworks of diagnosis offer substantial advantages, theoretically or practically, over the now well-established theories of diagnosis. In particular, it would be interesting to investigate the potentials of frameworks of diagnosis to act as a basis for building real-life applications.

Acknowledgements. I would like to thank Ameen Abu-Hanna, Frances Brazier, Luca Console, Linda van der Gaag, Frank van Harmelen, Niels Peek, Annette ten Teije, Jan Treur and two anonymous reviewers for their useful comments on earlier drafts of this article. I also would like to thank David Poole for making the Theorist system available to me.

References

[Allemang et al., 1987] D. Allemang, M.C. Tanner, T. Bylander and J. Josephson (1987). Computational complexity of hypothesis assembly. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, 1112–1117.

[Bankowitz et al., 1989] R.A. Bankowitz, M.A. McNeil, S.M. Challinor, R.C. Parker, W.N. Kapoor and R.A. Miller (1989). A computer-assisted medical diagnostic consultation service: implementation and prospective evaluation. *Annals of Internal Medicine*, **110**, 824–832.

[Beschta et al., 1993] A. Beschta, O. Dressler, H. Freitag, M. Montag, P. Struss (1993). DP-Net – a second generation expert system for localizing faults in power transmission networks. In *Proceedings of the International Conference on Fault Diagnosis* (Tooldiag 93), 1019–1027.

[Besnard, 1989] P. Besnard (1989). *An Introduction to Default Logic.* Berlin: Springer-Verlag.

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

[Brusoni et al., 1996] V. Brusoni, L. Console, P. Terenziani, D. Theseider Dupré (1996). Temporal model-based diagnosis: an overview from an abductive perspective. In *Proceedings of CESA'96 IMACS Multiconference: Symposium on Modelling, Analysis and Simulation*, **1**, 326–331.

[Buchanan & Shortliffe, 1984] B.G. Buchanan and E.H. Shortliffe (1984). *Rule-based Expert Systems: the MYCIN Experiments of the Stanford Heuristic Programming Project.* Reading: Addison-Wesley.

[Bylander et al., 1992] T. Bylander. D. Allemang, M.C. Tanner and J.R. Josephson (1992). The computational complexity of abduction. In *Knowledge Representation* (R.J. Brachman, H.J. Levesque and R. Reiter, eds.), 25–60. Cambridge, Massachusetts: The MIT Press.

[Campbell, 1976] E.J.M. Campbell (1976). Basic science, science and medical education. *Lancet*, **i**, 134–136.

[Charniak & Shimony, 1994] E. Charniak & S.E. Shimony (1994). Cost-based abduction and MAP explanation. *Artificial Intelligence*, **66**, 345–374.

[Clancey & Letsinger, 1984] W.J. Clancey and R. Letsinger (1984). NEOMYCIN: Reconfiguring a rule-based expert system for application to teaching. In *Readings in Medical Artificial Intelligence: the First Decade* (W.J. Clancey and E.H. Shortliffe, eds.). Reading, Massachusetts: Addison-Wesley.

[Clancey & Shortliffe, 1984] W.J. Clancey and E.H. Shortliffe, eds. (1984). *Readings in Medical Artificial Intelligence: the First Decade.* Reading, Massachusetts: Addison-Wesley.

[Clancey, 1985] W.J. Clancey (1985). Heuristic classification. *Artificial Intelligence*, **27**, 289–350.

[Clark, 1978] K.L. Clark (1978). Negation as failure. In *Logic and Databases* (H. Gallaire and J. Minker, eds.), 293–322. New York: Plenum Press.

[Console et al., 1989] L. Console, D. Theseider Dupré and P. Torasso (1989). A theory of diagnosis for incomplete causal models. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, 1311–1317.

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

[Console & Torasso, 1989] L. Console and P. Torasso (1989). A multi-level architecture for diagnostic problem solving. *Computational Intelligence*, **1**, 101–112.

[Console & Torasso, 1990a] L. Console and P. Torasso (1990). Hypothetical reasoning in causal models. *International Journal of Intelligent Systems*, **5**, 83–124.

[Console & Torasso, 1990b] L. Console and P. Torasso (1990). Integrating models of correct behaviour into abductive diagnosis. In *Proceedings of ECAI'90*, 160–166.

[Console & Torasso, 1991] L. Console and P. Torasso (1991). A spectrum of logical definitions of model-based diagnosis, *Computational Intelligence*, **7**(3), 133–141.

[Cox & Pietrzykowski, 1987] P.T. Cox and T. Pietrzykowski (1987). General diagnosis by abductive inference. In *Proceedings of the IEEE Symposium on Logic Programming*, 183–189.

[Dague, 1994] P. Dague (1994). Model-based diagnosis of analog electronic circuits. *Annals of Mathematics and Artificial Intelligence*, **11**, 439–492.

[Davis, 1983] R. Davis (1983). Reasoning from first principles in electronic trouble shooting. *International Journal of Man Machine Studies*, **19**(3), 403–424.

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

[Davis & Hamscher, 1988] R. Davis and W. Hamscher (1988). Model-based reasoning: troubleshooting. In *Exploring Artificial Intelligence: Survey Talks from the National Conference on Artificial Intelligence* (H.E. Shrobe, ed.), 297–346. San Mateo, California: Morgan Kaufmann.

[Davis & Shrobe, 1983] R. Davis and H. Shrobe (1983). Representing structure and behaviour of digital hardware. *IEEE Computer*, **16**(10), 75–82.

[Downing, 1993] K.L. Downing (1993). Physiological applications of consistency-based diagnosis. *Artificial Intelligence in Medicine*, **5**, 9–30.

[Duda et al., 1979] R.O. Duda, J. Gaschning and P.E. Hart (1979). Model design in the PROSPECTOR consultant program for mineral exploration. In *Expert Systems in the Microelectronic Age* (D. Michie, ed.). Edinburgh: Edinburgh University Press.

[Forbus & De Kleer, 1993] K.D. Forbus and J. de Kleer (1993). *Building Problem Solvers.* Cambridge, Massachusetts: The MIT Press.

[Fox et al., 1990b] J. Fox, C. Gordon, A.J. Glowinski and M. O'Neil (1990). Logic engineering for knowledge engineering: the Oxford System of Medicine. *Artificial Intelligence in Medicine*, **2**, 323–339.

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

[Genesereth & Nilsson, 1987] M.R. Genesereth and N.J. Nilsson (1987). *Logical Foundations of Artificial Intelligence.* Palo Alto, CA: Morgan Kaufmann.

[Hamscher, 1994] W. Hamscher (1994). CROSBY: financial data interpretation as model-based diagnosis. *Annals of Mathematics and Artificial Intelligence*, **11**, 511–524.

[Huang et al., 1993] J. Huang, J. Fox, C. Gordon and A. Jackson-Smale (1993). Symbolic decision support in medical care. *Artificial Intelligence in Medicine*, **5**(5), 415–430.

[Ishizuka et al., 1981] M. Ishizuka, K.S. Fu and J.T.P Yao (1981). Inexact inference for rule-based damage assessment of existing structures. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 837–842.

[Johnson & Keravnou, 1988] L. Johnson and E.T. Keravnou (1988). *Expert Systems Architectures.* London: Kogan Page.

[Josephson & Josephson, 1994] J.R. Josephson and S.G. Josephson (1994). *Abductive Inference: Computation, Philosophy, Technology.* Cambridge: Cambridge University Press.

[De Kleer, 1976] J. de Kleer (1976). Local methods for localizing faults in electronic circuits. *MIT AI Memo* 394. Cambridge, MA: Massachusetts Institute of Technology.

[De Kleer, 1977] J. de Kleer (1977). Multiple representation of knowledge in mechanic problem solving. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, 299–304.

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

[De Kleer & Williams, 1989] J. de Kleer and B.C. Williams (1989). Diagnosis with behavioural modes. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, 1324–1330.

[De Kleer et al., 1992] J. de Kleer, A.K. Mackworth and R. Reiter (1992). Characterizing diagnoses and systems. *Artificial Intelligence*, **52**, 197–222.

[Konolige, 1994] K. Konolige (1994). Using default and causal reasoning in diagnosis. *Annals of Mathematics and Artificial Intelligence*, **11**, 97–135.

[Kulikowski & Weis, 1982] C.A. Kulikowski and S.M. Weis (1982). Representation of expert knowledge for consultation: the CASNET and EXPERT projects. In *Artificial Intelligence in medicine* (P. Szolovits, ed.), 21–56. Boulder: Westview Press.

[Lewis & Papadimitriou, 1981] H.R. Lewis & C.H. Papadimitriou (1981). *Elements of the Theory of Computation.* Englewood Cliffs, NJ: Pretice-Hall.

[Lucas, 1993] P.J.F. Lucas (1993). The representation of medical reasoning models in resolution-based theorem provers. *Artificial Intelligence in Medicine*, **5**(5), 395–414.

[Lucas, 1996a] P.J.F. Lucas (1996). Modelling interactions for diagnosis. In *Proceedings of CESA'96 IMACS Multiconference: Symposium on Modelling, Analysis and Simulation*, **1**, 541–546.

[Lucas, 1996b] P.J.F. Lucas (1996). *A Theory of Diagnosis as Hypothesis Refinement.* Report UU-CS-1996-42. Utrecht: Department of Computer Science, Utrecht University.

[Lucas & Janssens, 1991] P.J.F. Lucas and A.R. Janssens (1991). Development and validation of HEPAR, an expert system for the diagnosis of disorders of the liver and biliary tract. *Medical Informatics*, **16**, 259–270.

[Lucas & Van der Gaag, 1991] P.J.F. Lucas and L.C. van der Gaag (1991). *Principles of Expert Systems.* Wokingham: Addison-Wesley.

[Macartney, 1988] F.J. Macartney (1988). Diagnostic logic. In *Logic in Medicine.* (C. Philips, ed.). London: British Medical Journal.

[McCarthy, 1986] J. McCarthy (1986). Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence*, **28**, 89–116.

[Miller et al., 1982] R.A. Miller, H.E. Pople and J.D. Myers (1982). INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, **307**, 468–476.

[Newell & Simon, 1972] A. Newell and H.A. Simon (1972). *Human Problem Solving.* Englewood Cliffs, NJ: Prentice-Hall.

[Ng, 1991] H.T. Ng (1991). Model-based, multiple fault diagnosis of dynamic, continuous physical devices. *IEEE Expert*, **6**(6), 38–43.

[Patil, 1981] R.S. Patil (1981). *Causal representation of Patient illness for electrolyte and acid-base diagnosis.* Technical Report MIT/LCS/TR-267. Cambridge, MA: Massachusetts Institute of Technology.

[Patil et al., 1982] R.S. Patil, P. Szolovits and W.B. Schwartz (1982). Modeling knowledge of the patient in acid-base and electrolyte disorders. In *Artificial Intelligence in Medicine* (P. Szolovits, ed.). Boulder, CO: Westview Press.

[Pearl, 1988] J. Pearl (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference.* San Mateo, CA: Morgan Kaufmann.

[Peng, 1985] Y. Peng (1985). *A Formalization of Parsimonious Covering and Probability Inference.* PhD thesis, Department of Computer Science, University of Maryland.

[Peng & Reggia, 1990] Y. Peng and J.A. Reggia (1990). *Abductive inference models for diagnostic problem solving.* New York: Springer-Verlag.

[Poole et al., 1987] D. Poole, R. Goebel and R. Aleliunas (1987). Theorist: a logical reasoning system for defaults and diagnosis. In *The Knowledge Frontier* (N. Cercone and G. Mc Calla, eds.), 331–352. Berlin: Springer-Verlag.

[Poole, 1988] D. Poole (1988). Representing knowledge for logic-based diagnosis. In *Proceedings of the International Conference on Fifth Generation Computer Systems 1988*, 1282–1290. ICOT.

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

[Poole, 1990a] D. Poole (1990). Normality and faults in logic-based diagnosis. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, 1304–1310.

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

[Poole, 1990c] D. Poole (1990). *A Theorist to Prolog Compiler.* Technical Report. Department of Computer Science, University of British Columbia.

[Poole, 1994] D. Poole (1994). Representing diagnosis knowledge. *Annals of Mathematics and Artificial Intelligence*, **11**, 33–50.

[Pople, 1973] H.E. Pople (1973). On the mechanization of abductive logic. In *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*, 147–152.

[Pople, 1977] H.E. Pople (1977). The formation of composite hypotheses in diagnostic problem solving: an exercise in synthetic reasoning. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, 1030–1037.

[Popper, 1959] K.R. Popper (1959). *The Logic of Scientific Discovery.* London: Hutchingson.

[Preist et al., 1994] C. Preist, K. Eshghi and B. Bertolino (1994). Consistency-based and abductive diagnosis as generalized stable models. *Annals of Mathematics and Artificial Intelligence*, **11**, 51–74.

[Punch III et al., 1990] W.F. Punch III, M.C. Tanner, J.R. Josephson and J.W. Smith (1990). PEIRCE: a tool for experimenting with abduction. *IEEE Expert*, **5**(5), 34–44.

[Reggia et al., 1983] J.A. Reggia, D.S. Nau and Y. Wang (1983). Diagnostic expert systems based on a set covering model. *International Journal of Man Machine Studies*, **19**, 437–460.

[Reiter, 1977] R. Reiter (1977). On closed world databases. In *Logic and Databases* (H. Gallaire and J. Minker, eds.). Berlin: Springer-Verlag.

[Reiter, 1980] R. Reiter (1980). A logic for default reasoning. *Artificial Intelligence*, **13**, 81–132.

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

[Santos, 1994] E. Santos Jr (1994). A linear constraint satisfaction approach to cost-based abduction. *Artificial Intelligence*, **65**, 1–27.

[Santos & Santos, 1996] E. Santos Jr and E.S. Santos (1996). Polynomial solvability of cost-based abduction. *Artificial Intelligence*, **86**, 157–170.

[Sauthier & Faltings, 1992] E. Sauthier and B. Faltings (1992). Model-based traffic control, *Artificial Intelligence in Engineering*, **7**, 139–151.

[Schaerf & Cadoli, 1995] M. Schaerf and M. Cadoli (1995). Tractable reasoning with approximation. *Artificial Intelligence*, **74**(2), 249–310.

[Shortliffe, 1976] E.H. Shortliffe (1976). *Computer-based Medical Consultations: MYCIN.* New York: Elsevier.

[Smith et al., 1985] J.W. Smith, J.R. Svirbely, C.A. Evans, P. Strohm, J.R. Josephson and M.C. Tanner (1985). RED: a red-cell antibody identification expert module. *Journal of Medical Systems*, **9**(3), 121–138.

[Stefanini et al., 1993] A. Stefanini, G. Tornielli and S. Cermignani (1993). An interval-based model for fault diagnosis in power transmission nets. In *Proceedings of the International Conference on Fault Diagnosis* (Tooldiag 93), 1000–1008.

[Struss, 1992] P. Struss (1992). What is in SD? Towards a theory of modelling for diagnosis. In *Readings in Model-based Diagnosis* (W. Hamscher, L. Console and J. de Kleer, eds.), 419–449. San Mateo: Morgan Kaufmann.

[Struss & Dressler, 1989] P. Struss and O. Dressler (1989). Physical negation: integrating fault models into the general diagnostic engine. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, 1318–1223.

[Szolovits, 1982] P. Szolovits (1982). *Artificial Intelligence in Medicine.* Boulder, CO: Westview Press.

[Ten Teije & Van Harmelen, 1994] A. ten Teije and F. van Harmelen (1994). An extended spectrum of logical definitions for diagnostic systems. In *DX-94, 5th International Workshop on Principles of Diagnosis* (G.M. Provan, ed.), 334–342.

[Ten Teije & Van Harmelen, 1996] A. ten Teije and F. van Harmelen (1996). Using approximate entailment for diagnostic reasoning. In *Proceedings of Principles of Knowledge Representation '96.*

[Torasso & Console, 1989] P. Torasso and L. Console (1989). *Diagnostic problem solving.* London: North Oxford Academic Publishers.

[Tuhrim et al., 1991] S. Tuhrim, J. Reggia and S. Goodall (1991). An experimental study of criteria for hypothesis plausibility. *Journal of Experimental and Theoretical Artificial Intelligence*, **3**, 129–144.

[Weiss et al., 1978] S.M. Weiss, C.A. Kulikowski, S. Amarel and A. Safir (1978). A model-based method for computer-aided medical decision making. *Artificial Intelligence*, **11**, 145–172.

[Wu, 1991] T.D. Wu (1991). A problem decomposition method for efficient diagnosis and interpretation of multiple disorders. *Computer Methods and Programs in Biomedicine*, **35**, 239–250.