ELSEVIER

Artificial Intelligence 105 (1998) 295–343

Artificial Intelligence

Analysis of notions of diagnosis

Peter J.F. Lucas *

Department of Computer Science, Utrecht University, P.O. Box 80.089, 3508 TB Utrecht, The Netherlands

Received 27 November 1997

Abstract

Various formal theories have been proposed in the literature to capture the notions of diagnosis underlying diagnostic programs. Examples of such notions are: heuristic classification, which is used in systems incorporating empirical knowledge, and model-based diagnosis, which is used in diagnostic systems based on detailed domain models. Typically, such domain models include knowledge of causal, structural, and functional interactions among modelled objects. In this paper, a new set-theoretical framework for the analysis of diagnosis is presented. Basically, the framework distinguishes between 'evidence functions', which characterize the net impact of knowledge bases for purposes of diagnosis, and 'notions of diagnosis', which define how evidence functions are to be used to map findings observed for a problem case to diagnostic solutions. This set-theoretical framework offers a simple, yet powerful tool for comparing existing notions of diagnosis, as well as for proposing new notions of diagnosis. A theory of flexible notions of diagnosis, called refinement diagnosis, is proposed and defined in terms of this framework. Relationships with notions of diagnosis known from the literature are investigated. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Diagnostic systems; Semantics of diagnosis; Formal theory of diagnosis

1. Introduction

Diagnostic computer programs were among the first systems developed in the field of applied Artificial Intelligence. In the burgeoning field of expert systems in the 1970s and 1980s, diagnostic applications abound. Although these systems frequently dealt with similar, or related, problem domains, often their underlying principles differed considerably. In a sense, this was a consequence of the additional goal of the development of many of these, now classic, programs: to explore representation and problem-solving methods. Only after researchers experienced that developing reliable diagnostic systems

* Email: lucas@cs.uu.nl.

PII: S0004-3702(98)00081-2

was much more difficult than previously thought, it was recognized that the principles underlying diagnosis were actually poorly understood.

Starting about halfway through the 1980s, a significant amount of research on conceptual and formal aspects of diagnosis was undertaken, with the aim of acquiring more insight into the nature of diagnostic problem solving. For example, Chandrasekaran and colleagues have analysed the diagnostic process conceptually in terms of a small number 'generic problem-solving tasks' [6]. Instead of studying the problem-solving behaviour of diagnostic systems, other researchers have focussed on representation issues in diagnostic systems. Where in many early diagnostic systems diagnostic knowledge from experts was captured in the form of empirical classification rules [4], in later systems model-based approaches became increasingly popular for building diagnostic systems in both industrial (cf. [2,14]), and nonindustrial areas, such as medicine (cf. [19,40]). The model-based approach advocates the construction of knowledge-based systems based on explicit models of a problem domain. For example, such models describe the structural and functional interactions among components of a physical system, or the causal interactions among elements in a domain. Model-based diagnosis was, in fact, already explored in the early systems INTER [17], SOPHIE [3], CASNET [43], and ABEL [27].

Although the introduction of the model-based approach to building diagnostic applications had a significant impact on the field of diagnosis, it did not immediately provide deep insight into the process of diagnosis. Real fundamental understanding of the nature of the diagnostic process was yielded by subsequent research concerning the formal aspects of diagnosis.

An early formal theory of diagnosis was proposed by Reggia and colleagues in terms of set theory, called *set-covering theory*, or *parsimonious covering theory* [35]. Basically, in the set-covering theory of diagnosis, causal knowledge of abnormality is represented by means of a binary causal relation, which is employed for computing diagnoses, essentially by determining whether actually observed findings can be predicted using the causal relation. Subsequent work has yielded several algorithms to compute set-covering diagnoses efficiently in practical applications [29,38,44], although this type of diagnostic reasoning is known to be NP-hard in general [5]. Experimental studies of set-covering theory and its variants have been performed by several researchers [21,34,41].

The formal aspects of diagnosis employing causal knowledge have also been studied, using logic as the primary tool [11,13,30,32]. In the logical theory of *abductive diagnosis*, diagnosis is formalized as reasoning from effects to causes, with causal knowledge represented as logical implications of the form

$$\mathit{causes} \rightarrow \mathit{effects}$$

where causes are usually abnormalities or faults, but they may also include normal situations. This abductive type of reasoning is contrasted with deduction, which for implications of the form above and under certain conditions, like that given *causes* and *effects* are conjunctions of positive literals, would amount to reasoning from causes to effects. Because in set-covering theory causal relations are also exploited to find causes for certain observed findings, this theory may be viewed as a specific theory of abductive diagnosis as well. The logical theory of abductive diagnosis is more expressive than set-covering theory, because it is possible to explicitly represent various types of interaction, which is not possible in the original set-covering theory. For example, it is not possible to express in the original set-covering theory that the simultaneous occurrence of two or more causes leads to the masking of certain findings. Console and colleagues have proposed several different versions of abductive diagnosis [9,11], and have also developed an implementation of the theory as the CHECK system [40]. Poole and colleagues have investigated abductive diagnosis using Theorist, a theory and system for hypothetical reasoning [30–32].

Approximately at the same time, Reiter proposed yet another logic-based theory of diagnosis, aiming at formally capturing diagnosis of abnormal behaviour in a device or system, using a model of normal structure and behaviour [36]. Nowadays, Reiter's theory, which was later extended by de Kleer and colleagues to deal with knowledge of both normal and abnormal behaviour [18], is usually referred to as the theory of *consistency-based diagnosis*. Basically, consistency-based diagnosis amounts to finding faulty device components that account for a discrepancy between predicted normal behaviour of a device, possibly supplemented with predictions of abnormal behaviour, both according to a domain model, and actually observed behaviour. The discrepancy is formalized as logical inconsistency; a diagnosis is established when assuming particular components to be faulty and others to be normally functioning restores consistency.

The abductive and consistency-based theories of diagnosis are often contrasted with diagnosis based on empirical associations. When empirical associations are represented as logical implications, then viewed as a classification relation, establishing a diagnosis can be accomplished by logical deduction, computing the closure of this classification relation. The more procedurally oriented term *heuristic classification* is frequently employed to refer to this type of diagnostic reasoning [8].

It has been shown that abductive and consistency-based diagnosis can be mapped to each other [36]. Furthermore, both types of diagnosis can be defined, in slightly different ways, in terms of the logical entailment relation [12,33]. Hence, although it was once thought that diagnostic systems could be classified as being either based on consistency checking, abductive reasoning, deductive reasoning, or on a combination of these three types of reasoning, it appears that characterizing diagnostic systems is more complicated than that [26].

The conclusion that there is not a unique way to characterize a particular type of diagnosis raises questions concerning the assumptions underlying abductive diagnosis, consistency-based diagnosis and heuristic classification. Does the logical notion of consistency provide an appropriate basis for formalizing various notions of diagnosis, and similarly, is logical implication the proper way to formalize relationships between causes and effects in abductive diagnosis, and to formalize empirical associations in heuristic classification? In this paper, it is argued that the formalization of diagnosis will benefit from the modelling of interactions at two levels of specification: (1) the modelling of interactions between the presence or absence of defects or faults, expressed by a mapping from defects to observable findings, and (2) the modelling of an interpretation of observed findings in the context of such a mapping. A set-theoretical semantic framework to express these aspects of diagnosis is proposed in Sections 2 and 3, and examined in detail in Section 4 using known theories of diagnosis from the literature. Medicine and simple logic circuits are taken as example domains.

As in many other theories of diagnosis, diagnostic problem solving is viewed as a special instance of *hypothetical reasoning* [31], possibly producing multiple, competing diagnoses. In solving a diagnostic problem, a hypothesis concerning the presence or absence of faults or abnormal processes, such as disorders in medicine, is first generated and next tested with respect to diagnostic knowledge and observed findings, yielding diagnoses that are best in a particular sense. However, since the result of this paper is a mathematical framework, no particular problem-solving order is enforced. The set-theoretical framework is expressive enough to go beyond common notions of diagnosis. This point is illustrated by the development of a theory of flexible diagnosis in Section 5, called *refinement diagnosis*, which is defined in terms of this framework. Relationships with notions of diagnosis known from the literature are investigated.

2. Interactions among defects and observables

There exists a strong relationship between the suitability of a particular type of knowledge for building a diagnostic system and the nature of the underlying problem domain. For example, for the construction of medical diagnostic systems, knowledge of the pathophysiology of disease processes can be used, but in other medical domains, like neurology, diagnostic problem solving mainly relies on the description of normal function, in combination with knowledge of the anatomical structure of the human body [24]. Similarly, in technical domains, knowledge of the structure of a device, supplemented with knowledge of how particular components are expected to behave normally or abnormally, can be used for the purpose of diagnosis [15]. Despite such differences, any knowledge base of a diagnostic system incorporates representations of meaningful interactions among defects (faults or disorders) and observable findings. We shall examine a number of typical examples of diagnostic knowledge bases to illustrate these points.

2.1. Motivating examples

Frequently employed types of knowledge encoded in diagnostic systems are causal knowledge, knowledge of structure and functional behaviour, and empirical associations. A small, but typical, example of each of these types of knowledge is presented below.

Causal interactions. Consider the following piece of medical knowledge: "influenza causes fever and infection of the trachea and bronchial tree (tracheobronchitis), which causes a sore throat; if the patient suffers from asthma, shortness of breath (dyspnoea) will occur as well". In Fig. 1(a), the directed-graph representation of the causal knowledge embodied in this medical description is depicted, where an arc denotes a cause-effect relationship. The medical meaning ascribed to the elements in the causal graph is indicated in Fig. 1(b). Using logic as our representation language, the figure may be assumed to correspond to a *causal specification* $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$, where $\Delta = \{d_1, d_2, d_3\}$ denotes a set of disorders, $\Phi = \{f_1, f_2, f_3\}$ denotes a set of observable findings, and $\mathcal{R}$ is a collection of rules in propositional logic:

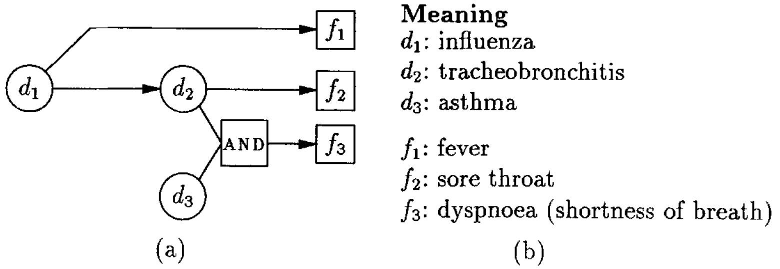


Fig. 1. Causal net.

$$\begin{aligned} d_1 &\rightarrow d_2 \\ d_1 &\rightarrow f_1 \\ d_2 &\rightarrow f_2 \\ d_2 \wedge d_3 &\rightarrow f_3. \end{aligned}$$

Such a causal specification is typically used in *abductive diagnosis* based on logic [9,32]. Note that the disorders d_1 and d_2 are causally related to each other; causal interaction is expressed by logical implication. A causal specification can be used for predicting observable findings. Assuming the presence of certain disorders, e.g., influenza (d_1), $\mathcal{R} \cup \{d_1\} \vDash \{f_1, f_2\}$ expresses that a patient with influenza will have symptoms and signs of fever (f_1) and sore throat (f_2) via a causal mechanism, where $\vDash$ denotes standard logical entailment. Here, the interaction between disorders, and between disorders and observable findings, is monotonic, due to the monotonic nature of ordinary logical entailment: by assuming more disorders, more observable findings will be predicted. Below, we shall consider various desirable nonmonotonic interactions, and also the qualitative representation of uncertain relationships between a cause and its associated effects.

Functional behaviour. Knowledge of normal and abnormal functional behaviour can be effective for diagnosing device problems, where the behaviour of the device is described in terms of relationships between input and output signals. These relationships are obtained from knowledge of the behaviour of the device's components and of the way in which these components are interconnected, i.e., the structure of the device. Consider a logic circuit consisting of an XOR (exclusive OR) gate X and an AND gate A, as shown in Fig. 2. The three input signals to the circuit are indicated by I_1, I_2 and I_3; O_1 and O_2 denote the two output signals.

Following the approach in [18], the normal behaviour of the two components can be described by the following two logical implications:

$$\forall x\big((\mathrm{XORG}(x) \wedge \neg\mathrm{Abnormal}(x)) \rightarrow \mathit{out}(x) = \mathit{xor}(\mathit{in}1(x), \mathit{in}2(x))\big)$$
$$\forall x\big((\mathrm{ANDG}(x) \wedge \neg\mathrm{Abnormal}(x)) \rightarrow \mathit{out}(x) = \mathit{and}(\mathit{in}1(x), \mathit{in}2(x))\big)$$

supplemented with the atoms XORG(X) and ANDG(A), which represent the XOR gate X and the AND gate A, respectively (uppercase symbols like X and A indicate constant

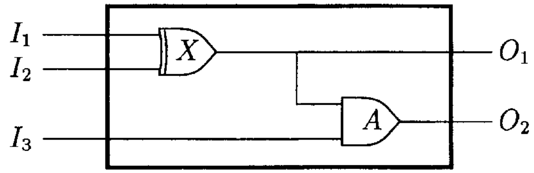


Fig. 2. Logic circuit.

symbols). If we assume that a defective gate always produces an output signal that is different from the correct output signal, the following logical specification of abnormal behaviour is obtained:

$$\forall x\big((\mathrm{XORG}(x) \wedge \mathrm{Abnormal}(x)) \rightarrow \mathit{out}(x) \neq \mathit{xor}(\mathit{in}1(x), \mathit{in}2(x))\big)$$
$$\forall x\big((\mathrm{ANDG}(x) \wedge \mathrm{Abnormal}(x)) \rightarrow \mathit{out}(x) \neq \mathit{and}(\mathit{in}1(x), \mathit{in}2(x))\big).$$

The structure of the circuit can be described by a collection of equalities, which also indicates how components influence each other, as follows:

$$\begin{array}{ll} \mathit{out}(X) = \mathit{in}1(A) & \mathit{in}2(A) = I_3 \\ \mathit{in}1(X) = I_1 & \mathit{out}(X) = O_1 \\ \mathit{in}2(X) = I_2 & \mathit{out}(A) = O_2. \end{array}$$

Now, a *system* $\mathcal{S}$ is defined as a pair $\mathcal{S} = (\mathrm{SD}, \mathrm{COMPS})$, consisting of a system description SD, such as the logical specification of the structure and behaviour of the circuit given above, with a set of components COMPS, with COMPS $= \{X, A\}$ in the present case. Such a specification is typically used in *consistency-based diagnosis*.

Suppose that the input to the circuit is as follows: $I_1 = 1$, $I_2 = 0$ and $I_3 = 1$. Using standard logical entailment, the following output can be predicted, assuming that the circuit is functioning correctly:

$$\mathrm{SD} \cup \{I_1 = 1, I_2 = 0, I_3 = 1, \neg\mathrm{Abnormal}(X), \neg\mathrm{Abnormal}(A)\}$$
$$\vDash \{O_1 = 1, O_2 = 1\}.$$

Similarly, partially abnormal behaviour, assuming part of the components to be abnormal, or completely abnormal behaviour can be predicted.

Empirical associations. Empirical associations represent knowledge derived from experience, and usually have the intended meaning of classification rules. Let $\mathcal{B} = (\Delta, \Phi, \mathcal{R}')$ denote an *associational specification* corresponding to the causal medical knowledge depicted in Fig. 1; the corresponding set of associational logical rules $\mathcal{R}'$ is defined as follows:

$$f_1 \rightarrow d_1$$
$$f_2 \rightarrow d_2$$
$$f_2 \wedge f_3 \rightarrow d_3$$

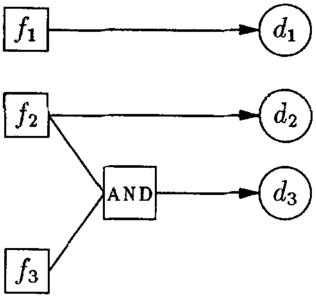


Fig. 3. Associations.

which can also be depicted as a directed graph, as shown in Fig. 3. Here, logical implication is interpreted as a classification relation; e.g., $\mathcal{R}' \cup \{f_2, f_3\} \vDash \{d_2, d_3\}$, i.e., sore throat and dyspnoea are classified as being due to the presence of tracheobronchitis together with asthma. Logical classification relations are typically used in heuristic classification.

Above, both the terms 'disorder' and 'fault' were used to refer to malfunction. Since the term 'disorder' is not used in technical domains, and the term 'fault' is not used in the biomedical domain, the term '*defect*' will be used henceforth to denote both disorders in medicine and faults in technical devices.

2.2. *The interpretation of knowledge as evidence*

As has been illustrated above, diagnostic systems may incorporate a wide variety of knowledge. In fact, in addition to the types of knowledge explored in the examples above, many other types of knowledge that are useful in a diagnostic setting can be distinguished. When building a particular diagnostic system, decisions concerning the type of knowledge to be included are clearly important. However, diagnostic systems also have a number of features in common; these features are particularly relevant when comparing systems. It appears that all diagnostic systems incorporate knowledge of interactions among defects and observable findings, which can be captured by means of particular mappings. These mappings, called 'evidence functions' in this paper, will be shown to offer a precise interpretation of the significance of the knowledge available to a system for the purpose of diagnosis, and will be introduced below.

Let Δ denote a countable set of *assumptions* concerning defects and possibly also concerning observable findings, as sometimes necessary for the representation of functional behaviour. Let Φ be a countable set of findings. To be able to make a distinction between present and absent defects and findings, respectively, a negation function $\neg$ is introduced. Positive defects d (findings f) and negative defects $\neg d$ (findings $\neg f$) denote *present* defects (findings) and *absent* defects (findings), respectively. It is assumed that the law of double negation holds, i.e., $\neg(\neg x) = x$, for a defect or finding x. If a defect or finding is not included in a set, it is assumed to be *unknown*. Let Δ_P (Φ_P) denote a set of positive defects (findings), and let Δ_N (Φ_N) denote a set of negative defects (findings), such that Δ_P (Φ_P) and Δ_N (Φ_N) are disjoint. It is assumed that $\Delta = \Delta_N \cup \Delta_P$ and $\Phi = \Phi_P \cup \Phi_N$. Often a set of assumptions Δ and a set of findings Φ will be disjoint, in which case Δ

merely consists of defects. To ease the exposition, this will be assumed in the following, unless stated otherwise. The power set of a set S is denoted by $\wp(S)$. As a matter of convenience, members of Δ_N are frequently denoted by $\neg d$, where $\neg(\neg d) = d \in \Delta_P$. Similarly, members of the set Φ_N are denoted by $\neg f$, where again $\neg(\neg f) = f \in \Phi_P$.

The intended diagnostic meaning of a knowledge base of a diagnostic system in terms of defects and findings is called a diagnostic specification. It is formally introduced below.

Definition 1 (*Diagnostic specification*). A *diagnostic specification* Σ is a triple $\Sigma = (\Delta, \Phi, e)$, where e is a function

$$e : \wp(\Delta) \to \wp(\Phi) \cup \{\perp\}$$

called an *evidence function*, such that:

(1) for each $f \in \Phi$ there exists a set $D \subseteq \Delta$ with $f \in e(D)$ or $\neg f \in e(D)$ (and possibly both);

(2) for each $D \subseteq \Delta$: if $d, \neg d \in D$ then $e(D) = \perp$;

(3) for each $D, D' \subseteq \Delta$: 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, and D is *consistent*; otherwise, it is said that D is *inconsistent*.

The set $e(D)$ stands for the set of observable findings for a set D of simultaneously occurring (present or absent) defects. In terms of diagnostic problem solving, the set $e(D)$ consists of findings that may be interpreted in some way as 'evidence' for the occurrence of the set of defects D, depending on the findings actually observed. How an evidence function may be used for diagnosing a problem is discussed in Section 3; in this section, we confine ourselves to investigating the meaning of evidence functions.

According to the definition above, we may have that both $f \in e(D)$ and $\neg f \in e(D)$, which simply means that these findings may alternatively occur given the combined occurrence of the defects in the set D, i.e., both f and $\neg f$ are associated with D. In some domains it might hold that if $e(\{d\}) = e(\{d'\})$, it follows that $d = d'$, i.e., the defects d and d' are taken as synonyms for the same defect. For example, if the defects stand for disorders in medicine, then two different names d and d' for which the equality holds, would normally be taken as different names for the same disorder. This situation is quite common in medicine. In general, sets of observable findings associated with defects may have several findings in common; thus, the sets $e(D)$ and $e(D')$, $D \neq D'$, need not be disjoint.

As indicated in Definition 1, a set of defects may be inconsistent just because it holds that $d, \neg d \in D$. This is a form of inconsistency that is evident for syntactic reasons. However, it is also possible that D is inconsistent for other than syntactic reasons, for example, because D contains defects d and d' that are incompatible. In this situation, the inconsistency is a consequence of a semantic relationship between the defects d and d'. In several definitions, it will be convenient to consider only sets of defects that are consistent for syntactic reasons; hence, the following definition:

Definition 2 (*Syntactic consistency*). Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, then the set of defects $D \subseteq \Delta$ is called *syntactically consistent* if for each defect $d \in D$: $\neg d \notin D$; otherwise, D is called *syntactically inconsistent*.

Next, the notion of maximal syntactic consistency is introduced; it is employed in the following to define particular evidence functions.

Definition 3 (*Maximal syntactic consistency*). Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, then the set of defects $D \subseteq \Delta$ is called *maximally syntactically consistent* if D is syntactically consistent and there exists no $d \in \Delta$, $d \notin D$, such that $D \cup \{d\}$ is syntactically consistent.

Sometimes, a knowledge base is only examined with respect to a *hypothesis* H, a subset of the entire set of defects Δ. For this purpose, the following definition is introduced.

Definition 4 (*Restricted evidence function*). Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification. A *restricted evidence function* of e with respect to a set $H \subseteq \Delta$, denoted by $e_{|H}$, is a function

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

such that for each $D \subseteq H$: $e_{|H}(D) = e(D)$.

From a general point of view, the expressive power of evidence functions is as large as infinite propositional logic; the function e may be viewed as similar to the conjunctive normal form of propositional formulae with defects and findings as literals. For example, the evidence-function representation of an implication $(d_1 \wedge d_2) \to (f_1 \vee f_2)$ would yield, among other function values, $e(\{d_1, d_2, \neg f_1\}) = \{f_2\}$. (Note that the argument $\{d_1, d_2, \neg f_1\}$ is allowed, because Δ and Φ need not be disjoint.) Hence, an evidence function is expressive enough to capture the sort of knowledge as represented in the logic theories of diagnosis, as introduced in Section 2.1. Consider the following example:

Example 5. Reconsider Fig. 1 and the associated logical specification of causal knowledge in Section 2.1. The following diagnostic specification $\Sigma = (\Delta, \Phi, e)$, where $\Delta_P = \{d_1, d_2, d_3\}$ and $\Phi_P = \{f_1, f_2, f_3\}$, corresponds to this causal specification. The intended meaning of this causal specification with respect to diagnosis can be captured by means of an evidence function e for which, among others, the following holds:

$$\begin{array}{llll}
e(\{d_1\}) & = \{f_1, f_2\} & e(\{d_3\}) & = \emptyset \\
e(\{d_2\}) & = \{f_2\} & e(\{d_2, d_3\}) & = \{f_2, f_3\} \\
e(\{d_1, d_2\}) & = e(\{d_1\}) & e(\{d_1, d_2, d_3\}) & = e(\{d_1, d_3\}) = \{f_1, f_2, f_3\} \\
e(\{d_1, \neg d_2, d_3\}) & = \bot. & &
\end{array}$$

The property $e(\{d_i\}) \subseteq e(\{d_1, d_2\})$, $i = 1, 2$, formally expresses that the interaction between d_1 and d_2 is monotonic; the evidence function e is monotonically increasing.

The value $e(\{d_1, \neg d_2, d_3\}) = \bot$ indicates an impossible situation, because if d_1 is present, then d_2 cannot be absent (though, it may be unknown); $\{d_1, \neg d_2, d_3\}$ is inconsistent for semantical reasons. The evidence function e actually extends the logic specification in Section 2.1, by assuming that the specification is also intended to deal with negative defects.

The reader has probably noticed that the evidence function above can be specified more tersely; in Section 2.4 techniques for the partial specification of evidence functions will be discussed in detail.

For a diagnostic system incorporating knowledge of structure and of normal or abnormal behaviour, the following diagnostic specification is obtained.

Example 6. Reconsider the logic circuit depicted in Fig. 2, with the associated system $\mathcal{S}$ provided in Section 2.1. Suppose that presence of a defect in X is denoted by x; absence of a defect in X is denoted by $\neg x$. A similar notation is employed with respect to gate A. If $I_j = 1$, this will be denoted by i_j; an input equal to $I_j = 0$ will be denoted by $\neg i_j$. A similar convention is adopted for the output signals O_k. Fixed input signals to the circuit are $i_1, \neg i_2$ and i_3. Now, the output signals are represented as observable findings, and a component for which the presence or absence of a defect is unknown, is taken into account by assuming that the component is either defective or nondefective. The following evidence function (only values for consistent sets of defects are provided) corresponds to the description above:

$$\begin{aligned}
e'(\{x, a\}) &= \{\neg o_1, o_2\} \\
e'(\{\neg x, a\}) &= \{o_1, \neg o_2\} \\
e'(\{x, \neg a\}) &= \{\neg o_1, \neg o_2\} \\
e'(\{\neg x, \neg a\}) &= \{o_1, o_2\} \\
e'(\{x\}) &= \{\neg o_1, o_2, \neg o_2\} \\
e'(\{\neg x\}) &= \{o_1, o_2, \neg o_2\} \\
e'(\{a\}) &= \{o_1, \neg o_1, o_2, \neg o_2\} \\
&= e'(\{\neg a\}) \\
&= e'(\emptyset).
\end{aligned}$$

For example, $e'(\{x\}) = \{\neg o_1, o_2, \neg o_2\}$ indicates that when the XOR gate X is defective, and it is unknown whether or not the AND gate A is defective, then the first output signal $O_1 = 0$ and the second output signal O_2 may be either 0 or 1, depending on whether the AND gate is defective or not. Hence, $e'(\{x\})$ is defined with respect to the output of the entire circuit in Fig. 2, not merely the output produced by the output channel directly connected to the XOR gate, i.e., O_1. For this circuit in general, the observable findings for $e'(D)$ always include o_1, $\neg o_1$, or both, and o_2, $\neg o_2$, or both. In contrast with the assumptions underlying the evidence function given in Example 5, the behaviour of the system is described with respect to all elements of the entire system, and not in terms of isolated (defective) components.

The two examples above were meant to convey some intuition concerning the expressive power of evidence functions for capturing the semantical significance of knowledge for the purpose of diagnosis. One of the attractive features of evidence functions is that they provide an easy means for describing properties of diagnostic interpretations of knowledge bases in a precise, formal way.

2.3. *Properties of evidence functions*

As has been argued above, an evidence function e may possess certain properties, determined by the (diagnostic) knowledge incorporated in the knowledge base on which it is based. In this section, an overview is provided of properties of evidence functions that will be useful for characterizing diagnostic knowledge. Some of these properties will be required in the analysis of the various formal theories of diagnosis in Section 4.

The various properties can be distinguished into *global* properties, i.e., properties that hold for the entire evidence function e, and *local* properties, i.e., properties that hold only for some sets of defects D.

2.3.1. *Global properties*

In descriptions of many problem domains, only positive findings, or positive findings and a few negative findings, are employed to characterize sets of defects. This situation has already been encountered in Example 5. By the definition of an evidence function (cf. Definition 1), any finding that is included in the set of findings Φ must appear either positively, negatively, or both, in some function value $e(D)$, $D \subseteq \Delta$. This explains why from Definition 1 it follows that

$$\bigcup_{D \subseteq \Delta, D \text{ consistent}} e(D) = \Phi$$

need not hold, because for some finding $f \in \Phi$, we may have that $\neg f \notin e(D)$, for each $D \subseteq \Delta$. Nevertheless, sometimes every positive *and* negative finding in Φ is covered by the evidence function e. The consequence is that such an evidence function is, in principle, dependent on the notion of diagnosis employed, capable of producing a diagnosis for any set of findings observed (cf. Section 3).

Monotonicity of an evidence function is a property that will be encountered several times in the analysis of theories of diagnosis in Section 4. It is defined as follows:

Definition 7 (*Monotonicity*). Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification. The evidence function e is called *monotonically increasing* if

$$\forall D, D' \subseteq \Delta \colon\ D \subseteq D' \Rightarrow e(D) \subseteq e(D')$$

and e is called *monotonically decreasing* if

$$\forall D, D' \subseteq \Delta \colon\ D \subseteq D' \Rightarrow e(D) \supseteq e(D')$$

with D and D' consistent. If e is either monotonically increasing or decreasing, it is called *monotonic*; otherwise, e is called *nonmonotonic*.

If an evidence function is monotonically increasing, this means that the more defects are considered, the more (new) findings must be taken into account. The evidence function in Example 5, which was the result of the translation of causal knowledge into evidence-function representation, was monotonically increasing. If an evidence function is monotonically decreasing, this means that if more defects are considered, information concerning the observable findings of sets of defects will be more specific. We have encountered an example of such a function in Example 6, where knowledge concerning the normal and abnormal behaviour of a circuit was encoded. Note that what is often referred to as the 'nonmonotonicity of diagnosis' (cf. [36]) actually concerns the interpretation of observed findings in the process of diagnosis. This is an aspect completely different from the one considered in this section, but will be considered in Section 3.

Of special interest in the previous section was the representation of interactions among defects and findings in terms of an evidence function. If no interactions among defects and findings exist (except inconsistency among syntactically inconsistent defects), the evidence function conforms to the following definition:

Definition 8 (*Interaction freeness*). A set of defects Δ of a diagnostic specification $\Sigma = (\Delta, \Phi, e)$ is called *interaction free with respect to e* if

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

for each syntactically consistent set of defects $D \subseteq \Delta$. If in addition for each $d \in \Delta$: $e(\{d\})$ is nonempty, and for each $d, d' \in \Delta$, $d \neq d'$, it holds that

$$e(\{d\}) \cap e(\{d'\}) = \emptyset$$

the set Δ is called *strongly interaction free*; otherwise, Δ is called *weakly interaction free*.

We will sometimes simply say that the evidence function e is interaction free. Interaction freeness means that the observable findings associated with a collection of defects D are the same as the collected observable findings associated with each individual defect $d \in D$. Thus, by combining the observable findings for individual defects, the observable findings for combinations of defects are obtained. Although interaction freeness is presented here as a global property, we shall occasionally employ the phrase in a *local* sense, to express that two or more defects do not interact with each other, e.g., $e(\{d, d'\}) = e(\{d\}) \cup e(\{d'\})$. It is easy to show that an evidence function that is interaction free is also monotonically increasing.

Proposition 9. *If $\Sigma = (\Delta, \Phi, e)$ is a diagnostic specification, such that Δ is interaction free, then e is monotonically increasing.*

Proof. Simply note that if $D \subseteq D'$, with consistent sets $D, D' \subseteq \Delta$, then

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

From this, it follows that $e(D) \subseteq e(D')$. □

As a matter of convenience, function values $e(\{d\})$ of an evidence function that defines Δ to be interaction free, are sometimes simply denoted by $e(d)$. If a set of defects is strongly interaction free with respect to some evidence function e, this does not necessarily imply that the defects do not influence each other in one way or the other; it only means that these influences have not been captured in the function e explicitly, because the meaning attached to e does make these influences irrelevant with respect to diagnosis.

In some domains in which defects are interaction free, it holds that each defect is described in unique terms, i.e., for each defect $d \in \Delta$, the set of observable findings $e(d)$ is not contained in the set $e(D)$, if d is not included in D. It is shown that the evidence function restricted to consistent sets of defects is injective (the notation $V \backslash W$ stands for the difference between the two sets V and W).

Proposition 10. *Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification such that Δ is interaction free with respect to e. Then, if for each $d \in \Delta$, and each consistent set $D \subseteq \Delta \backslash \{d\}$, it holds that $e(\{d\}) \not\subseteq e(D)$, then the restriction of the evidence functions e to consistent subsets of Δ is injective.*

Proof. It has to be proven that for consistent $D, D' \subseteq \Delta$, with $D \neq D'$, it holds that $e(D) \neq e(D')$. If $D \neq D'$, then there exists a defect $d \in D$ (or $d \in D'$ if $D \subset D'$, but reversing D and D' does not matter), such that $d \notin D'$. Hence, according to the assumption of the proposition: $e(d) \not\subseteq e(D')$. Since it holds by interaction freeness that $e(d) \subseteq e(D)$, it follows, also from interaction freeness, that $e(D) \not\subseteq e(D')$. From this, the result follows immediately. □

Given this proposition, the following corollary holds:

Corollary 11. *If $\Sigma = (\Delta, \Phi, e)$ is a diagnostic specification such that Δ is strongly interaction free, then the restriction of the evidence function e to syntactically consistent sets of defects is injective.*

Proof. Simply note that if Δ is strongly interaction free, it holds that $e(\{d\}) \not\subseteq e(D \backslash \{d\})$ for each $D \subseteq \Delta$. □

Proposition 10 is also satisfied if for each $d \in \Delta$, $e(d)$ includes a unique observable finding (called a *pathognomonic* finding in medicine). Note that it is now possible to uniquely identify a set of defects D by its associated set of observable findings $F = e(D)$, due to the injective nature of e (but the set of defects may also be undefined). This yields a very simple form of diagnosis.

2.3.2. *Local properties*

There are a number of local properties of evidence functions that are the result of mapping a semantic relationship among defects to relationships among sets of observable findings. A typical example of such a relationship is causality. For example, if the defect d is known to cause the defect d', it is, in terms of the associated evidence function, known that the set of observable findings for d contains all observable findings associated with d', i.e.,

$$e(\{d'\}) \subseteq e(\{d\}). \tag{1}$$

In a monotonic theory of causality (cf. [9,10]), the following would hold as well:

$$e(\{d\}) = e(\{d, d'\}) \tag{2}$$

expressing that as d causes d', when d and d' are present together, precisely the same set of observable findings would be obtained as if only d was present and d' is unknown. From (1) and (2) it follows that

$$e(\{d, d'\}) = e(\{d\}) \cup e(\{d'\}).$$

Hence, d and d' are assumed to be interaction free in the local sense; note that d and d' are only weakly interaction free. This is not a global property of causality as employed in abductive diagnosis, because interaction freeness will not hold in general (cf. Example 5).

Note that a causal specification C with a set of logical rules equal to

$$\mathcal{R} = \{d_1 \rightarrow d_2,\ d_2 \rightarrow f\}$$

is not distinguishable in terms of evidence functions from

$$\mathcal{R}' = \{d_1 \leftrightarrow d_2,\ d_2 \leftrightarrow f\}$$

because in both cases an interaction-free evidence function e with $e(\{d_i\}) = \{f\}$, $i = 1, 2$, results. This means that $\mathcal{R}$ and $\mathcal{R}'$ are similar with respect to their diagnostic interpretation.

Starting with causality in a more general sense, a number of local properties of evidence functions will be examined.

(a) *Influence interactions*: the occurrence of some defects influences the occurrence of other defects, as reflected by the observable findings. The following two types of local interaction are distinguished:

- *Causality*: if the combination of defects D causes the set of findings F, then $F = e(D)$. The diagnostic view of knowledge of the sort 'the set of defects D causes the set of defects D'' as, for example, used in abductive diagnosis can be made precise in terms of an evidence function as follows:

  $$e(D') \subseteq e(D)$$

  for some consistent $D, D' \subseteq \Delta$, i.e., any finding that may be observed for the set of defects D' may also be observed for the set of defects D. Furthermore, in this case it holds that

  $$e(D) = e(D \cup D').$$

  In Example 5 above, simple causal relationships between three individual defects were examined.

  Various other types of causal relations can be expressed in terms of evidence functions. For example, the values of the evidence function

  $$e(\{d_1\}) = e(\{\neg d_i\}) = \emptyset$$

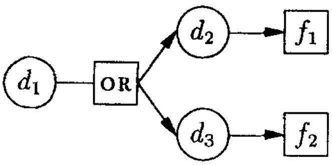


Fig. 4. Nondeterministic causality.

for $i = 1, \ldots, 3$, and

$$e(\{d_2\}) = e(\{d_1, \neg d_3\}) = \{f_1\}$$
$$e(\{d_3\}) = e(\{d_1, \neg d_2\}) = \{f_2\}$$

express *nondeterministic causality* between the defect d_1 on the one hand, and d_2 and d_3, on the other hand, as depicted in Fig. 4.

- *Correlation*: if the defects d and d', $d \neq d'$, are correlated, then if d has occurred then d' occurs as well, and vice versa, whereas if d is absent ($\neg d$), d' is also absent ($\neg d'$), and vice versa. Correlation of defects can be described by means of an evidence function as follows:

$$\begin{aligned} &e(\{d\}) &&= e(\{d'\}) &&= e(\{d, d'\}) \\ &e(\{\neg d\}) &&= e(\{\neg d'\}) &&= e(\{\neg d, \neg d'\}) \\ &e(\{d, \neg d'\}) = e(\{\neg d, d'\}) = \bot. \end{aligned}$$

The conditions above are satisfied for positive correlation; negative correlation can be described by means of the conditions

$$\begin{aligned} &e(\{d\}) &&= e(\{\neg d'\}) = e(\{d, \neg d'\}) \\ &e(\{\neg d\}) &&= e(\{d'\}) \quad = e(\{\neg d, d'\}) \\ &e(\{d, d'\}) = \bot. \end{aligned}$$

(b) *Synonyms*: if the defects $d \in \Delta$ and $d' \in \Delta$ are synonymous, then $e(\{d\}) = e(\{d'\})$. This is commonly applied in medicine, as has been discussed above. If for each $d, d' \in \Delta$, $d \neq d'$, it holds that $e(\{d\}) \neq e(\{d'\})$, there are no synonymous defects. It is said that Δ (also e) is *synonym free*.

(c) *Synergic interactions*: these are interactions that augment, cancel, preclude, exclude, or complement local interactions among defects. The following types of interaction are distinguished:

- *Augmentation* (also referred to as *potentiation*): the combined occurrence of two or more defects in the set D gives rise to new observable findings in addition to those associated with the individual elements, or proper subsets of D, i.e.,

$$e(D) \supset \bigcup_{D' \subset D} e(D') \tag{3}$$

for some consistent $D \subseteq \Delta$. It is interesting to note that (3) is yielded for monotonically increasing evidence functions, using the weaker condition:

$$e(D) \not\subseteq \bigcup_{D' \subset D} e(D').$$

- *Cancellation* (also referred to as *fault masking* [15] or *antagonism*): the combined occurrence of two or more defects in the set D yields fewer observable finding when compared to the findings associated with the individual elements, or proper subsets of D, i.e.,

$$e(D) \subset \bigcup_{D' \subset D} e(D')$$

for some consistent $D \subseteq \Delta$.

- *Augmented cancellation*: this notion combines the notions of augmentation and cancellation mentioned above, after weakening both conditions. The following holds:

$$e(D) \not\subseteq \bigcup_{D' \subset D} e(D') \wedge e(D) \not\supseteq \bigcup_{D' \subset D} e(D')$$

for some consistent $D \subseteq \Delta$. For example, $e(\{d_1\}) = \{f_1\}$, $e(\{d_2\}) = \{f_2, f_3\}$, but $e(\{d_1, d_2\}) = \{f_3, f_4\}$; hence, the findings f_1 and f_2 are cancelled, and a new finding (f_4) is observable. Note that $e(\{d_1, d_2\}) \circ e(d_i)$, $i = 1, 2$, fails to hold for $\circ \in \{\subset, \supset\}$. This is a consequence of the dependence between augmentation and cancellation. The cancellation of findings causes augmentation to fail, and vice versa. Hence, the weakening of the two conditions in the notion of augmented cancellation.

- *Preclusion*: the presence of one or more defects in a combination implies that each element in some other combination of defects is assumed to be absent. This can be expressed by:

$$e(\{d_1, \ldots, d_n\}) \supseteq e(\{\neg d_1', \ldots, \neg d_m'\}).$$

This means that a set of present defects contains information pertaining to a set of absent defects. Note that if Δ is interaction free, it follows that

$$e(\{d_1, \ldots, d_n\}) \supseteq e(\{\neg d_i'\})$$

for each i, $1 \leqslant i \leqslant m$, $m \geqslant 1$. This yields a preclusion relation that is more easy to grasp, namely that a combination of defects D precludes some defect d:

$$e(D) \supseteq e(\neg d).$$

- *Exclusion*: some combination of defects D cannot occur:

$$e(D) = \bot.$$

- *Complementation*: the observable findings associated with the absent defects $\neg d_1, \ldots, \neg d_n$, are the complements of those associated with the presence of those, i.e., if $e(\{d_1, \ldots, d_n\}) = \{f_1, \ldots, f_m\}$ then $e(\{\neg d_1, \ldots, \neg d_n\}) = \{\neg f_1, \ldots, \neg f_m\}$.

(d) *Empirical associations*: when the defects in the set D are simultaneously present, the findings in the set F may be observed, given $F = e(D)$. Knowledge based on empirical associations is often structured according to individual defects and families (categories) of defects; a defect d can be called *more specific* than a defect d' if $e(\{d\}) \subset e(\{d'\})$; if this relation holds for more than one defect d, then defect d' may be taken as a *category* (it includes a number of different defects).

The evidence-function representations of causal knowledge and of empirical associations have much in common, but there are a few differences. Firstly, the condition $e(d) = e(d')$ fails to hold for empirical associations if d and d' are not synonymous. Secondly, a defect d for which $e(d) \supset e(d')$, for more than one defect $d' \in \Delta$, will be a category if the evidence function e stands for empirical associations, but, d will not be a category in general if e represents causal knowledge.

This concludes our list of various interactions among defects, and their expression in terms of evidence functions.

2.4. *Partial specification*

When a domain satisfies certain properties, it may be sufficient to provide a partial specification of an evidence function. Partial specification has the virtue that it is not always necessary to explicitly specify, or compute, the exponential number of function values of an evidence function e; it suffices to provide only part of them explicitly. Any algorithm for diagnosis using an evidence function of the form discussed in the previous section, without simplifying assumptions, will be intractable. In [5], in which the complexity of algorithms for abductive diagnosis is analysed, it is therefore assumed that the specification of a domain theory is polynomial in the sum of the cardinalities of the sets Δ and Φ. A *partial specification* of an evidence function e consists of a restriction of e, denoted by $\tilde{e}$, which is defined on a nonempty subset $V \subseteq \wp(\Delta)$, together with a number of computation rules expressing how function values $e(D)$ must be determined. If an evidence function is defined by means of a partial specification, it is called *partially specified*.

In domains for which not all function values $e(D)$ can be provided explicitly, such as in medicine, the condition that the specification of an evidence function is polynomial in size is usually fulfilled, be it for pragmatic reasons. In biomedical applications there is usually insufficient knowledge available to explicitly capture all interactions among defects, because the medical literature provides little information about the observable features of specific disorder combinations. In technical applications, the situation is less unfavourable, in the sense that often precise technical descriptions of the domain are available.

In several diagnostic theories, for example, the set-covering theory of diagnosis [29], partial specification includes a restriction of an evidence function to singleton sets, i.e., it suffices to define an evidence function in terms of the individual defects distinguished in the domain. If the associated computation rule expresses that the observable findings for nonsingleton sets of defects can be taken as the union of the observable findings associated with their elements, the evidence function is interaction free. This limitation is enforced by some formal theories of diagnosis; it may not be sanctioned by the characteristics of a problem domain, as we have seen in the previous section.

Although the extension of a partial specification to an evidence function is thus dependent on known evidence-function properties, expressed by means of computation rules, there are two extremes that deserve attention. The first useful way of partially specifying an evidence function is based on the assumption that when no explicit knowledge concerning the findings associated with a set of defects D is available, implicitly the largest proper subsets D' of D for which $\tilde{e}(D')$ is given, are taken to yield sufficient information concerning the interactions among the elements of D. This form of partial specification is called bottom-up partial specification.

Definition 12 (*Bottom-up partial specification*). Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, and let $V \subseteq \wp(\Delta)\backslash\{\emptyset\}$ be a set, such that for each $d \in \Delta$: $\{d\} \in V$. Then, a function

$$\tilde{e} : V \to \wp(\Phi) \cup \{\bot\}$$

is called a *bottom-up partial specification* of e if:

(1) for each $D \in V$: $e(D) = \tilde{e}(D)$;
(2) for each $D \in \wp(\Delta)\backslash V$:

$$e(D) = \bigcup_{\substack{D' \subset D,\ D' \in V \\ \forall D'' \in V,\ D'' \subset D:\ D'' \not\supset D'}} e(D').$$

Hence, by a bottom-up partial specification $\tilde{e}$ we mean a restriction of an evidence function e with appropriate computation rules to generate the function e from $\tilde{e}$. The principal idea of condition (2) is illustrated in Fig. 5, where a node in the graph represents a set and an edge represents proper set inclusion; all nodes below the node labelled D in the graph are proper subsets of D. Note that a restriction $\tilde{e}$ need not be unique; one can freely include subsets D of Δ in the domain of the restriction $\tilde{e}$ for which $e(D)$ could also be determined using condition (2) in the definition above. The intuitive idea of a bottom-up partial specification is that information concerning the interaction among defects is derived from the largest (with respect to $\subset$) proper subsets D' of a set of defects D, for which function values $\tilde{e}(D')$ have explicitly been given; the function value $e(D)$, when not explicitly given by $\tilde{e}$, is obtained as the union of all such $\tilde{e}(D')$. In the examples below this choice will be further clarified. For convenience, in the following, function values for syntactically inconsistent sets will be left out from the definition of bottom-up

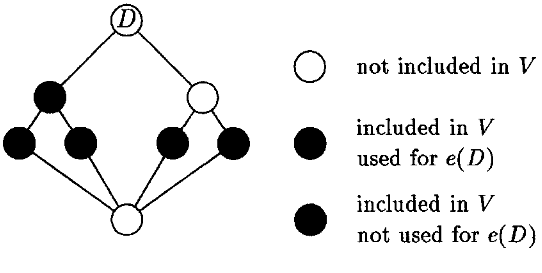

D
○ not included in V
● included in V used for e(D)
● included in V not used for e(D)

Fig. 5. Part of a lattice used for bottom-up specification of an evidence function.

partial specifications $\tilde{e}$. From the definition of a bottom-up partial specification it follows that $e(\emptyset) = \emptyset$, i.e., there are no observable findings if there is no knowledge concerning defects. If the problem domain concerns the (faulty) behaviour of a device, a bottom-up partial specification amounts to specifying the isolated behaviours of parts of the device. Hence, a bottom-up partial specification is in line with a specification of causal knowledge as in the abductive theory of diagnosis, i.e., any diagnostic specification obtained from this theory can be described as a bottom-up partial specification.

Often, the causal relation, such as represented by standard logical entailment, is taken to be monotonic. Bottom-up partial specifications, however, also allow for representing nonmonotonic interactions and complementary findings representing alternative observable findings, e.g., f and $\neg f$, thus extending the repertoire of the types of knowledge that can be used for diagnosis.

Example 13. 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 signs and symptoms 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. Fig. 6(a) depicts this simple problem as a directed graph; the meaning of the nodes in the graph is indicated in Fig. 6(b).

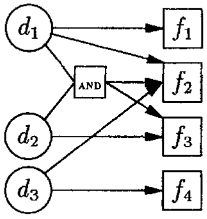
Consider a diagnostic specification $\Sigma = (\Delta, \Phi, e)$, where e is bottom-up partially specified by means of the function $\tilde{e}$, which is defined as follows:

$$\tilde{e}(D) = \begin{cases} \{f_1, f_2\} & \text{if } D = \{d_1\}, \\ \{f_3\} & \text{if } D = \{d_2\}, \\ \{f_2, f_4\} & \text{if } D = \{d_3\}, \\ \{f_2, f_3\} & \text{if } D = \{d_1, d_2\}, \\ \emptyset & \text{if } D = \{\neg d_i\}, i = 1, \ldots, 3. \end{cases}$$

From this specification, it follows that $e(\{d_1\}) \not\subseteq e(\{d_1, d_2\})$; e is nonmonotonic. Note the difference between Fig. 6, which is a representation of the function $\tilde{e}$, and does not assume monotonicity, and the evidence-function interpretation of Fig. 1. Here, it does not hold that $e(\{d_1, d_2\}) = \{f_1, f_2, f_3\}$.

As a prerequisite for bottom-up partial specification, it is assumed that at least knowledge concerning individual defects (i.e., singleton sets of defects) is available in a given diagnostic domain. This is not an unrealistic assumption, because in many problem domains knowledge concerning the possible abnormal behaviour resulting from an individual defect is the kind of knowledge most readily available.

Example 14. Consider again the evidence function from the example above (Example 13). From this partial specification it follows that, for example, $e(\Delta_P) = \tilde{e}(\{d_1, d_2\}) \cup \tilde{e}(\{d_3\}) =$



(a)

Meaning
d_1: pulmonary infection
d_2: Cushing's disease
d_3: iron-deficiency anaemia
f_1: fever
f_2: dyspnoea (shortness of breath)
f_3: moon face
f_4: low serum iron

(b)

Fig. 6. Partial evidence function $\tilde{e}$.

$\{f_2, f_3, f_4\}$, where $\Delta_P = \{d_1, d_2, d_3\}$. Note that neither $\tilde{e}(\{d_1\})$ nor $\tilde{e}(\{d_2\})$ play a role in determining $e(\Delta_P)$, because there is information available about the interaction between the defects d_1 and d_2 by the function value $\tilde{e}(\{d_1, d_2\})$. This function value provides partial information about the mutual influences among the defects in Δ_P; more precise information about the possible interactions between the members of Δ_P is unavailable; hence, $\{d_1, d_2\}$ and $\{d_3\}$ are assumed to be free of interaction, but the defects d_1 and d_2 are not.

It follows that a bottom-up partial specification may provide information about the interaction between defects. In the extreme situation that no interaction between defects exists, it suffices to define a partial specification in terms of individual defects only.

Proposition 15. *If $\Sigma = (\Delta, \Phi, e)$ is a diagnostic specification, such that Δ is interaction free, then there exists a bottom-up partial specification $\tilde{e}$ of e with domain*

$$V = \big\{\{d\} \mid d \in \Delta\big\}.$$

Proof. Note that if the domain of $\tilde{e}$, V, is defined as above, conditions (1) and (2) in Definition 12 simplify to the definition of interaction freeness; hence, the evidence function can be defined as follows

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

for each syntactically consistent set $D \subseteq \Delta$. □

The second typical form of a partial specification of an evidence function is obtained by providing at least explicit function values for maximally syntactically consistent sets $D \subset \Delta$, and describing other combinations of defects D' by taking associated observable findings of defects $d \notin D$ into account. In the following example, this particular partial specification technique is introduced, using a diagnostic description of a logic circuit taken from [18].

Example 16. Consider the logic circuit depicted in Fig. 7, which consists of two NOT gates (or inverters) in series. In [18], the problem of diagnosing faulty behaviour of the

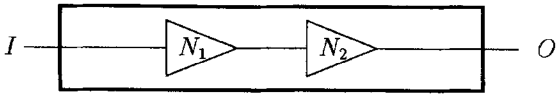


Fig. 7. Two NOT gates in series.

given logic circuit is described for an input signal fixed to $I = 0$, denoted here by $\neg i$, with resulting output signals equal to $O = 0$, denoted by $\neg o$, or $O = 1$, denoted by o, respectively. Again, output signals correspond to observable findings. The following behavioural assumptions are made in [18]. If a NOT gate N_i is defective, denoted by n_i, its output will be either 0, or the input to the gate is shorted (unmodified) to its output; $\neg n_i$ designates that the NOT gate N_i is not defective. Given this information, the following restriction $\tilde{e}$ of e of a diagnostic specification $\Sigma = (\Delta, \Phi, e)$ can be defined (we have disregarded the input, because it is assumed to be fixed to 0):

$$\begin{aligned}\tilde{e}(\{n_1, n_2\}) &= \{\neg o\}\\ \tilde{e}(\{\neg n_1, n_2\}) &= \{\neg o, o\}\\ \tilde{e}(\{n_1, \neg n_2\}) &= \{o\}\\ \tilde{e}(\{\neg n_1, \neg n_2\}) &= \{\neg o\}.\end{aligned}$$

The complementary pair $\{\neg o, o\}$ is the result of the assumption above that there are two different, nondeterministic types of abnormal behaviour. The function $\tilde{e}$ is taken as a partial specification to generate e by assuming that $e(\{n_1\}) = \{\neg o, o\}$, etc., meaning that if it is unknown whether or not N_2 is defective, the possible output of the circuit, given N_1 to be defective, is $\{\neg o, o\}$. Thus, similar to Example 6, we have that

$$e(\{n_1\}) = e(\{n_1, n_2\}) \cup e(\{n_1, \neg n_2\}).$$

Interestingly, this partial specification indicates that if the observed output signal is equal to o, either $\{\neg n_1, n_2\}$ or $\{n_1, \neg n_2\}$ may be the case, which are precisely the diagnostic alternatives provided by de Kleer et al. However, it is not at all obvious from their example that for an output equal to $\neg o$, the set of defects $\{\neg n_1, n_2\}$ is a possibility as well. This information is immediately available from the evidence function e.

This way of partially specifying an evidence function will be called top-down partial specification of an evidence function. A top-down partial specification is appropriate when it is not possible to describe defects with associated observable findings in isolation from other defects and associated findings, i.e., knowledge of the associated findings of the other defects, including their interaction, is needed to describe the defects. If the domain is a device, this assumption means that it is not possible to describe the (normal or abnormal) behaviour of a component in isolation from its environment. One could view the approach supported by top-down specification as a 'holistic approach', and the approach supported by bottom-up specification as a 'reductionistic approach'. Top–down partial specification is defined below.

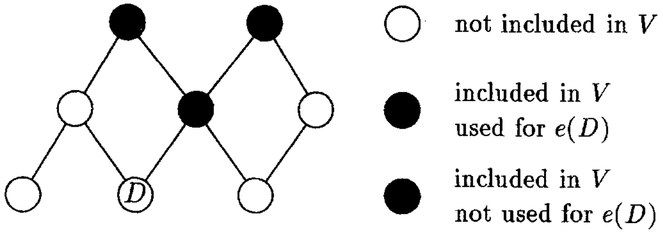


Fig. 8. Part of a lattice used for top-down specification of an evidence function.

Definition 17 (*Top-down partial specification*). Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification, and let $V \subseteq \wp(\Delta)\backslash\{\emptyset\}$ be a set, such that for each maximally syntactically consistent set $D \subseteq \Delta$: $D \in V$. Then, the function

$$\tilde{e} : V \to \wp(\Phi) \cup \{\bot\}$$

is called a *top-down partial specification* of e if:

(1) for each $D \in V$: $e(D) = \tilde{e}(D)$;

(2) for each $D \in \wp(\Delta)\backslash V$:

$$e(D) = \bigcup_{\substack{D' \supset D,\ D'\text{consistent},\ D' \in V\\ \forall D'' \in V,\ D'' \supset D:\ D'' \not\subset D'}} e(D').$$

Note that $e(D)$ is obtained by taking the union of all function values $e(D')$, where $D' \in V$ is a minimal proper superset of D, and no set $D'' \in V$ is smaller than D'. The principal idea is depicted in Fig. 8. In Examples 6 and 16, the behaviour of two logic circuits was studied using evidence functions that could have been generated by a top-down partial specification $\tilde{e}$, with

$$V = \bigl\{\{a, x\}, \{\neg a, x\}, \{a, \neg x\}, \{\neg a, \neg x\}\bigr\}$$

for Example 6 and

$$V = \bigl\{\{n_1, n_2\}, \{\neg n_1, n_2\}, \{n_1, \neg n_2\}, \{\neg n_1, \neg n_2\}\bigr\}$$

for Example 16. The assumption underlying an evidence function defined in this way is that it is sufficient to describe a domain in terms of the observable findings associated with all maximally consistent combinations of defects in the domain. This means that if the domain is a device consisting of components that may be defective, information about the isolated behaviour of individual components of the system has not been supplied. If a set of defects is described in terms of this special case of a top-down partial specification, we shall say that it is externally described.

Definition 18 (*Externally described*). Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification. The set of defects Δ is called *externally described* with respect to e if there exists a top-down partial specification $\tilde{e}$ for e with domain V, where for each $D \in V$: D is maximally syntactically consistent.

Note that if Δ is externally described with respect to e, the definition of the evidence function can be simplified as follows. For each consistent $D \subseteq \Delta$:

$$e(D) = \bigcup_{\substack{D' \supseteq D,\ D' \in V \\ D' \text{consistent}}} \tilde{e}(D').$$

It is easily shown that an evidence function for a set of defects that is externally described is monotonically decreasing.

Proposition 19. *If $\Sigma = (\Delta, \Phi, e)$ is a diagnostic specification, such that Δ is externally described, then e is monotonically decreasing.*

Proof. If $D \subseteq D'$, with consistent $D, D' \subseteq \Delta$, then

$$e(D') = e(D \cup D') = \bigcup_{\substack{D'' \supseteq (D \cup D'),\ D'' \in V \\ D'' \text{consistent}}} e(D'') \subseteq \bigcup_{\substack{D' \supseteq D,\ D' \in V \\ D' \text{consistent}}} e(D') = e(D).$$

From this, it follows that e is monotonically decreasing. □

Observe that top-down partial specification does not result in a significant reduction in the number of values to be specified for an evidence function, because if $|\Delta_P| = n$, at least 2^n function values have to be specified.

Above, we have introduced two opposite ways to define evidence functions. Bottom-up partial specification appeared to be particularly suitable for generating evidence functions for defects among which a limited amount of interaction exists. By contrast, top-down partial specification is most suitable for generating evidence functions for defects which are strongly interrelated. As one would expect, there are also evidence functions that lie somewhere between these two extremes, suitable for representing particular real-world knowledge.

3. Notions of diagnosis

As has been discussed, an evidence function can be viewed as a semantic interpretation of a knowledge base, containing, for example, causal or functional knowledge, in terms of expected evidence for the combined occurrence of (present or absent) defects. 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 observed findings that is adopted, can be viewed as a notion of diagnosis applied for solving the diagnostic problem at hand.

More formally, let $\mathcal{P} = (\Sigma, E)$ be a *diagnostic problem*, where $E \subseteq \Phi$ is a set of *observed findings*; it is assumed that if $f \in E$ then $\neg f \notin E$, i.e., contradictory observed findings are not allowed. The set of observed findings E denotes findings that are present or absent at a given time. In contrast, the findings in the set of observable findings $F = e(D)$,

$D \subseteq \Delta$, need not all be observed at the same time. Let R_Σ denote a *notion of diagnosis* R defined for suitable diagnostic specifications, and here applied to Σ, then a mapping

$$R_{\Sigma,e_{|H}} : \wp(\Phi) \to \wp(\Delta) \cup \{u\}$$

called a *diagnostic function*, will either provide a diagnostic solution for a diagnostic problem $\mathcal{P}$, or indicate that no solution exists, denoted by u (undefined). Recall that H denotes a *hypothesis*, which is taken to be a set of defects (more generally, assumptions $H \subseteq \Delta$), and $e_{|H}$ denotes the restricted evidence function of e. A notion of diagnosis R is usually a partial function; it is only defined for diagnostic specifications satisfying certain requirements.

Next, a diagnosis is defined as the result of applying a diagnostic function to a set of observed findings.

Definition 20 (*Diagnostic solution*). Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem, with $E \subseteq \Phi$ a set of observed findings. Let R be a notion of diagnosis. An *R-diagnostic solution*, or *R-diagnosis* for short, with respect to the set of defects $H \subseteq \Delta$ is defined as follows:

$$R_{\Sigma,e_{|H}}(E).$$

In Fig. 9, the idea underlying the definition of a notion of diagnosis R and diagnostic solution to a diagnostic problem is illustrated schematically.

The definition above is very unrestrictive; one reasonable restriction on the notion of diagnosis is obtained by assuming that for each nonempty $E \subseteq \Phi$, and each nonempty, consistent set $H \subseteq \Delta$, for which $R_{\Sigma,e_{|H}}(E) = H'$, with $H' \neq u$, it holds that $e_{|H}(H') \cap E \neq \emptyset$ if $e_{|H}(H') \neq \emptyset$, i.e., at least one observed finding in E must be accounted for by the diagnosis H'. The set of findings $e_{|H}(H') \cap E$ is called the set of findings *accounted for* by H'. The condition that at least one finding must be accounted for simply means that the result H' obtained by applying a diagnostic function has at least some relevance

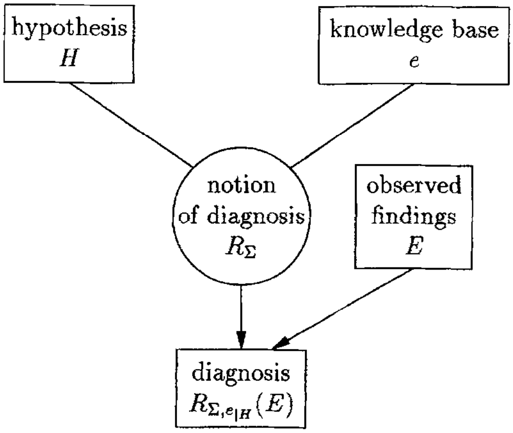


Fig. 9. Schema of notion of diagnosis, diagnostic problem and solution.

with respect to the findings observed. This is a rather weak condition. Other, more precise constraints will be encountered below for specific notions of diagnosis.

If application of the diagnostic function $R_{\Sigma,e_{|H}}$ yields as a result $H' = R_{\Sigma,e_{|H}}(E)$, it is said that:

(1) the hypothesis H is *accepted* if $H' = H$;
(2) the hypothesis H is *rejected* if $H' = u$;
(3) otherwise, the hypothesis H is said to be *adjusted.*

Adjustment of a hypothesis indicates that not all defects in H have passed when the hypothesis was tested against E, i.e., the result H' is taken as the adjusted version of the original hypothesis H.

Note that it is possible that

$$R_{\Sigma,e_{|H}}(E) = R_{\Sigma,e_{|H'}}(E)$$

for $H \neq H'$.

Example 21. To demonstrate how the definitions above can be employed, consider a notion of diagnosis U, such that $U_{\Sigma,e_{|H}}(E) = H'$ if it holds that H' is the largest subset of H with $e_{|H}(H') \subseteq E$; otherwise, $H' = u$. This notion of diagnosis expresses that a diagnosis consists of a set of defects which, on the one hand, can account for at least part of all observed findings, and, on the other hand, every finding associated with the set of defects that is taken as a diagnosis has been observed. Furthermore, there is only one such maximal subset of the given hypothesis H. Now, reconsider the medical domain from Example 13 (Fig. 6) with $H = \{d_1, d_2\}$ (pulmonary infection and Cushing's disease). Some interesting diagnostic conclusions are:

$$U_{\Sigma,e_{|H}}\big(\{f_1, f_2\}\big) = \{d_1\},$$

i.e., a patient with only fever and dyspnoea has pulmonary infection, $U_{\Sigma,e_{|H}}(\{f_1\}) = u$, i.e., there exists no diagnosis accounting for only fever as sign, and finally,

$$U_{\Sigma,e_{|H}}\big(\{f_2, f_3\}\big) = H.$$

In the first case, the hypotheses has been *adjusted,* in the second case, the hypothesis H is *rejected,* and in the last case, the hypothesis H has been *accepted.*

This example demonstrates the flexibility of the approach.

As remarked above, Definition 20 imposes very few constraints with respect to the properties that must be satisfied by a reasonable notion of diagnosis. One conceivable property that, however, usually fails to hold, is that a notion of diagnosis respects the evidence function e.

Definition 22 (*R respects e*). Let R be a notion of diagnosis defined for the diagnostic specification $\Sigma = (\Delta, \Phi, e)$. It is said that R *respects* e if:

(1) for each set of observed findings $E \subseteq \Phi$, there exists a set $H \subseteq \Delta$ such that $e(R_{\Sigma,e_{|H}}(E)) = E$, and
(2) for each consistent $D \subseteq \Delta$, there exists a set $H \subseteq \Delta$, such that $R_{\Sigma,e_{|H}}(e(D)) = D$ and for each $H' \not\supseteq H$: $R_{\Sigma,e_{|H'}}(e(D)) = u$.

This means that a function that is taken as the inverse of the evidence function e, which must be bijective (excluding inconsistent sets of defects and sets $E \subseteq \Phi$ with complementary findings), is composed of function values $R_{\Sigma,e_{|H}}(E)$, where the set $H \subseteq \Delta$ need not be fixed. Of course, the two conditions above will also hold if there exists a function $R_{\Sigma,e_{|H}}$ with fixed H that can be taken as the inverse.

If a notion of diagnosis respects an evidence function, and, in addition, an evidence function is interaction free, the following proposition holds:

Proposition 23. *If R is a notion of diagnosis defined for the diagnostic specification $\Sigma = (\Delta, \Phi, e)$, where e is interaction free, and R respects e, then*

$$R_{\Sigma,e_{|H}}(E) = R_{\Sigma,e_{|H}}(E') \cup R_{\Sigma,e_{|H}}(E'')$$

for each set of observed findings E, E' and E'', with $E, E', E'' \subseteq \Phi$ and $E = E' \cup E''$, and $H \subseteq \Delta$.

Proof. Since e is bijective if restricted to consistent sets of defects D, we know that there exist sets D, D' and D'' such that $E = e(D)$, $E' = e(D')$ and $E'' = e(D'')$, with $E = E' \cup E''$. Then, using the fact that e is interaction free: $e(D) = e(D') \cup e(D'') = e(D' \cup D'')$. Therefore, $D = D' \cup D''$, because e is injective. From the fact that R respects e it follows that

$$\begin{aligned} R_{\Sigma,e_{|H'}}(E') \cup R_{\Sigma,e_{|H''}}(E'') &= R_{\Sigma,e_{|H'}}(e(D')) \cup R_{\Sigma,e_{|H''}}(e(D'')) \\ &= D' \cup D'' \\ &= R_{\Sigma,e_{|H}}(e(D' \cup D'')) \\ &= R_{\Sigma,e_{|H}}(E) \end{aligned}$$

for some consistent $H, H', H'' \subseteq \Delta$. Furthermore, since R respects e and $R_{\Sigma,e_{|H}}(E) = D$ it follows that $e_{|H}(D) = E$ ($D \subseteq H$ holds by definition). Similarly, from $R_{\Sigma,e_{|H'}}(E') = D'$ we have $e_{|H'}(D') = E'$. Moreover, because $D' \subseteq D$ it follows that $D' \subseteq H$, hence $e_{|H}(D') = E'$. Therefore, $R_{\Sigma,e_{|H'}}(E') = R_{\Sigma,e_{|H}}(E')$. Analogously, $R_{\Sigma,e_{|H''}}(E'') = R_{\Sigma,e_{|H}}(E'')$. □

Hence, it turns out that if a notion of diagnosis R respects an interaction-free evidence function e, the set of observed findings can be partitioned, such that each subset can be accounted for separately by the same function $R_{\Sigma,e_{|H}}$. Note that if we have an evidence function e for which $f, \neg f \in e(D)$, for some $D \subseteq \Delta$, then R cannot respect e, due to the fact that E cannot contain complementary findings, at least, if $R_{\Sigma,e_{|H}}(E)$ is to be interpreted as a diagnosis.

A notion of diagnosis R provides the possibility to express interactions among defects and observed findings at the level of diagnosis, which we call dependencies. We may also have that a hypothesis can be split up into two subhypotheses, that can be examined independently, yielding a form of compositionality. More formally, we have the following definition:

Definition 24 (*Independence assumption*). Let R be a notion of diagnosis. It is said that R fulfills the *independence assumption* if for each diagnostic specification Σ for which

R_Σ is defined, and for each pair of consistent sets of defects $H, H' \subseteq \Delta$ and each set of observed findings $E \subseteq \Phi$ it holds that

$$R_{\Sigma,e_{|H\cup H'}}(E) = R_{\Sigma,e_{|H}}(E) \cup R_{\Sigma,e_{|H'}}(E)$$

with $R_{\Sigma,e_{|H\cup H'}}(E) \neq u$.

This means that the diagnostic solution with respect to the hypothesis $H \cup H'$ is obtained as the union of the solutions for the two separately examined hypotheses H and H'. As we shall see, for many notions of diagnosis described in the literature, in particular for abductive diagnosis and consistency-based diagnosis, the independence assumption fails to hold.

Example 25. The following notion of diagnosis S is defined for diagnostic specifications $\Sigma = (\Delta, \Phi, e)$, where the evidence function e is interaction-free. Let $E \subseteq \Phi$ be a set of observed findings, then

$$S_{\Sigma,e_{|H}}(E) = \bigcup_{H' \subseteq H,\, e_{|H}(H') \subseteq E} H'$$

for each consistent $H \subseteq \Delta$. The intuitive idea underlying this notion of diagnosis is that only defects in a hypothesis H that have all their associated findings included as observed findings are admitted as part of a diagnosis; the least upper bound of accepted subhypotheses is taken as the most likely diagnosis. The independence assumption is satisfied for S, because any interaction-free evidence function is monotonically increasing, therefore, if $e(D) \subseteq E$, then $e(D') \subseteq E$, $D' \subseteq D$.

Next, diagnostic monotonicity is defined for a notion of R-diagnosis; it is a property in line with the independence assumption.

Definition 26 (*Diagnostic monotonicity*). A notion of diagnosis R is called *diagnostically monotonic* if for each diagnostic specification Σ for which R_Σ is defined, each consistent set of defects $H \subseteq H'$, with $H, H' \subseteq \Delta$, and each set of observed findings $E \subseteq \Phi$, it holds that if $R_{\Sigma,e_{|H}}(E) \neq u$, then $R_{\Sigma,e_{|H}}(E) \subseteq R_{\Sigma,e_{|H'}}(E)$; otherwise, R is called *diagnostically nonmonotonic*.

Diagnostic monotonicity of a notion of diagnosis means: the larger (with respect to $\subseteq$) the hypothesis investigated, the larger the diagnostic solution. Note that from diagnostic monotonicity, it follows that if $H \subseteq H'$, then $e(R_{\Sigma,e_{|H}}(E)) \subseteq e(R_{\Sigma,e_{|H'}}(E))$ if e is monotonically increasing.

The following proposition states that any notion of diagnosis satisfying the independence assumption is diagnostically monotonic.

Proposition 27. *A notion of diagnosis R is diagnostically monotonic if the independence assumption is satisfied.*

Proof. Let R be a notion of diagnosis, then for every diagnostic specification Σ: if $H \subseteq H'$, with consistent $H, H' \subseteq \Delta$, and $R_{\Sigma,e_{|H}}(E) \neq u$, then $R_{\Sigma,e_{|H}}(E) \subseteq R_{\Sigma,e_{|H'}}(E)$, because $R_{\Sigma,e_{|H'}}(E) = R_{\Sigma,e_{|H}}(E) \cup R_{\Sigma,e_{|H' \setminus H}}(E)$. □

Independence and diagnostic monotonicity were introduced as properties of abductive diagnosis for the first time in [5].

In the next two sections, various notions of diagnosis are compared, and their diagnostic characteristics are explored. The two orderings defined below, will be employed frequently in such comparisons.

Definition 28 (*Restriction*). Let R and R' be two notions of diagnosis. Then, R is called a *restriction* of R', denoted by

$$R \sqsubseteq R'$$

if for each Σ, and for each $H \subseteq \Delta$, and set of observed findings $E \subseteq \Phi$ it holds that: if $R_{\Sigma,e_{|H}}(E) = H'$, $H' \neq u$, then $R'_{\Sigma,e_{|H}}(E) = H'$.

Thus, if the restriction relation between two notions of diagnosis R and R' holds, then the diagnoses resulting from the notion of diagnosis R are a subset of those resulting from R' (for any legal diagnostic specification Σ).

The notion of subdiagnostic relation is useful for characterizing the relative strictness in admitting defects to a diagnostic solution from notions of diagnosis.

Definition 29 (*Subdiagnostic relation*). Let R and R' be two notions of diagnosis. The notion of diagnosis R is called *subdiagnostic* to R', denoted by

$$R \trianglelefteq R'$$

if $R_{\Sigma,e_{|H}}(E) \subseteq R'_{\Sigma,e_{|H}}(E)$ given that $R_{\Sigma,e_{|H}}(E), R'_{\Sigma,e_{|H}}(E) \neq u$, for each Σ, and for each $H \subseteq \Delta$ and set of observed findings $E \subseteq \Phi$.

We shall occasionally employ the same symbol $\trianglelefteq$ to denote that the diagnostic solutions of some diagnostic function are a subset of those of another diagnostic function applied to the same diagnostic specification, i.e.,

$$R_{\Sigma,e_{|H}} \trianglelefteq R'_{\Sigma,e_{|H'}}$$

iff $R_{\Sigma,e_{|H}}(E) \subseteq R'_{\Sigma,e_{|H'}}(E)$, for each set of observed findings E, given that the diagnoses are defined.

4. Analysis of notions of diagnosis from the literature

Because the diagnostic formalism introduced above is meant to act as a framework, various notions of diagnosis known from the literature should be expressible in it. In this section, the expressive power of the framework is examined with respect to abductive and consistency-based diagnosis. Notions of diagnosis related to heuristic classification will be examined in the next section. Some obvious properties of notions of diagnosis shall be stated without proof (cf. [25] for complete proofs).

4.1. Abductive diagnosis

The formalization of diagnosis using causal domain models, usually referred to as *abductive diagnosis*, has been studied by several researchers [9,12,22,30–32]. A typical example is the work by Console and colleagues [9,12]. In their theory of abductive diagnosis, the abnormal or normal behaviour of a system is modelled in terms of causal knowledge with abnormal or normal states (called defects in this paper) and predicted findings as basic ingredients. Two different types of causal knowledge are distinguished in this theory. In the first type of causal knowledge, it is assumed that when a collection of defects is present, all causally associated findings *must* be present as well. This notion of causality will be called *strong causality*. In the second type of causal knowledge, causally related findings *may* be present, but need not be, when the associated defects are present. This less strict notion of causality will be called *weak causality*; it represents an imprecise uncertain relationship between cause and effect. We start by analysing diagnostic problem solving based on strongly causal knowledge, and next consider the usage of weakly causal knowledge and the consequences of combining both types of knowledge.

Strongly causal relationships among defects, and between defects and observable findings, are denoted in the theory of abductive diagnosis by logical implications of the form

$$d_1 \wedge \cdots \wedge d_n \to d$$

and

$$d_1 \wedge \cdots \wedge d_n \to f$$

expressing that the combined occurrence of defects $d_1, \ldots, d_n$ *causes* defect d and finding f, respectively, to occur; findings f and defects d are represented as ground literals in predicate logic. A *causal specification* $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$, which was already informally introduced in Section 2.1, is defined as a set of defect literals Δ, finding literals Φ, and a collection of logical implications $\mathcal{R}$ concerning defect and finding literals of the form above. The logical implications in $\mathcal{R}$ are often referred to as *abnormality axioms*, since they usually represent causal knowledge of abnormality only.

Now, let $\mathcal{A} = (\mathcal{C}, E)$ be an *abductive diagnostic problem*, with $\mathcal{C}$ a causal specification, and let E be a set of observed findings. Then, a set of defects $H \subseteq \Delta$ is called a *diagnosis* of $\mathcal{A}$ iff [9]:

(1) $\forall f \in E$: $\mathcal{R} \cup H \vDash f$ (*covering condition*), and

(2) $\forall f \in E^c$: $\mathcal{R} \cup H \nvDash \neg f$ (*consistency condition*)

where E^c, the set of observable findings assumed to be absent, is defined in terms of E as follows:

$$E^c = \{\neg f \in \Phi \mid f \notin E, f \text{ is a positive literal}\}.$$

This means that a diagnosis H must predict all findings observed, but may not predict findings assumed to be absent.

Example 30. Reconsider the causal specification $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$ from Section 2.1, as depicted in Fig. 1. Suppose that $E = \{f_1, f_2\}$ (fever and sore throat) is a given set of observed findings, then we have that $E^c = \{\neg f_3\}$ (dyspnoea is absent), and, thus, $H = \{d_1, d_2\}$ is a diagnosis for $\mathcal{A}$, because the covering and consistency conditions are satisfied.

This form of abductive diagnosis can be translated into our framework in a straightforward fashion. Let $\mathcal{P}$ represent the diagnostic problem corresponding to the abductive problem, such that $\mathcal{P} = \tau(\mathcal{A})$, where τ maps an abductive diagnostic problem $\mathcal{A}$ to a diagnostic problem $\mathcal{P}$ in our framework. To distinguish between elements of an abductive diagnostic problem $\mathcal{A}$ and a diagnostic problem $\mathcal{P}$, subscripts $\mathcal{A}$ and $\mathcal{P}$ will be attached to elements. The meaning of a causal specification $\mathcal{C}$ of an abductive diagnostic problem $\mathcal{A}$ is captured by an evidence function e with domain $\wp(\Delta_{\mathcal{P}})$ as follows. For each $D_{\mathcal{A}} \subseteq \Delta_{\mathcal{A}}$:

(1) if $\mathcal{R} \cup D_{\mathcal{A}}$ is satisfiable, then $e(D_{\mathcal{P}}) = \{\tau(f) \mid \mathcal{R} \cup D_{\mathcal{A}} \vDash f, f \in \Phi_{\mathcal{A}}\}$;

(2) otherwise, $e(D_{\mathcal{P}}) = \perp$,

where $D_{\mathcal{P}} = \tau(D_{\mathcal{A}})$. Condition (1) interprets causal knowledge in terms of predicting observable findings.

For ease of exposition, in the following, defects $\tau(d) \in \Delta_{\mathcal{P}}$ and defect literals $d \in \Delta_{\mathcal{A}}$ will not explicitly be distinguished; similarly, no difference is made between findings $\tau(f) \in \Phi_{\mathcal{P}}$ and finding literals $f \in \Phi_{\mathcal{A}}$.

Example 31. For the axioms $\mathcal{R}$ in the example above, the evidence function e of the corresponding diagnostic specification $\Sigma = (\Delta, \Phi, e)$ is given in Example 5.

Abductive diagnosis as defined above in terms of the covering and consistency conditions can now be defined as a notion of diagnosis. The corresponding notion of diagnosis is called the notion of *strong-causality diagnosis* (SC). It is defined as follows:

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

i.e., it is necessary that all observable findings $e(H)$ are observed (consistency condition), and vice versa (covering condition), to accept an hypothesis H as a diagnosis. This is just expressed by means of equality in our framework.

Example 32. For the diagnostic problem $\mathcal{P} = (\Sigma, E)$, with diagnostic specification Σ as in Example 31 and set of observed findings $E = \{f_1, f_2\}$, it is concluded that the patient has influenza and tracheobronchitis:

$$\mathrm{SC}_{\Sigma, e_{|\{d_1, d_2\}}}(\{f_1, f_2\}) = \{d_1, d_2\}$$

which is exactly the same result as obtained by abductive diagnosis in Example 30. Note that for $E' = \{f_1\}$ no abductive diagnosis exists. Similarly, it holds that $\mathrm{SC}_{\Sigma, e_{|H}}(E') = u$ for $E' = \{f_1\}$ and every consistent $H \subseteq \Delta$.

The notion of strong-causality diagnosis is not diagnostically monotonic, because it may hold that $\mathrm{SC}_{\Sigma,e_{|H}}(E) = H$, but $\mathrm{SC}_{\Sigma,e_{|H'}}(E) = u$, for $H \subset H'$ (which is easily shown by means of a counterexample). The independence assumption also fails to hold for strong-causality diagnosis (just take $H' \subset H$, with $H = \{d_1, d_2\}$ in the example above).

A notion of *weak* causality [9] is arrived at by the addition of *assumption literals* α to the individual abnormality axioms. These literals are employed to express that a causal relation is uncertain. Hence, the abnormality axioms $\mathcal{R}$ of an abductive diagnostic problem $\mathcal{A}$ are of one of the following two forms:

$$d_1 \wedge \cdots \wedge d_n \wedge \alpha_d \to d$$
$$d_1 \wedge \cdots \wedge d_n \wedge \alpha_f \to f$$

expressing that the combined occurrence of defects $d_1, \ldots, d_n$ *may* cause defect d and finding f, respectively, to occur. The transformation τ introduced above must be extended in order to deal with the assumption literals expressing weak causality. There are two possibilities. First, the abnormality axioms $\mathcal{R}$ could be translated to an evidence function e, where the assumption literals in a solution H are taken as defects, i.e., if for $f \in E$

$$\mathcal{R} \cup H \vDash f$$

and $\mathcal{R} \cup H$ is satisfiable, then $f \in e(H)$, where H is a set of defects, possibly including assumption literals α, i.e., $d = \tau'(\alpha)$, with transformation τ' extending τ, and d a defect. Next, the notion of diagnosis SC introduced above for strong causality could be employed as diagnostic interpretation of the resulting evidence function e. Obviously, weak causality is then expressed at the level of the knowledge base, i.e., at the object-level. The second possibility amounts to lifting the notion of weak causality to the meta-level: a notion of diagnosis is designed that interprets a knowledge base containing causal knowledge as being weakly causal in nature. Let A denote the set of assumption literals in $\Delta_{\mathcal{A}}$. Then, $\mathcal{R}'$ is a set of abnormality axioms obtained by removing each assumption literal $\alpha \in A$ from the axioms in $\mathcal{R}$. The transformation τ'' is then defined in the same way as τ, except that $\mathcal{R}'$ replaces $\mathcal{R}$.

The theory of abductive diagnosis adopts the first approach, because the same covering and consistency conditions are employed to define diagnosis for weakly causal knowledge as for strongly causal knowledge. Here, the second approach may also be adopted, i.e., uncertainty in causal knowledge is lifted to the level of diagnostic interpretation.

To study the difference in diagnostic interpretation of evidence functions with respect to weak and strong causality, a distinction is made between an abductive *solution*—a set of defects and assumption literals for which the covering and consistency conditions are satisfied—and an (abductive) *diagnosis*, the set of all defects included in an abductive solution.

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:

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

A weak-causality diagnosis accounts for all observed findings, although not every (predicted) observable finding need be observed.

Note that for the notions of diagnosis SC and WC, it holds that

$$e(R_{\Sigma,e_{|H}}(E)) \supseteq E$$

if $R_{\Sigma,e_{|H}}(E) \neq u$, where $R \in \{\mathrm{SC}, \mathrm{WC}\}$, i.e., all findings that have actually been observed must have been predicted as being observable for the associated set of defects. This is a consequence of the fact that for any abductive diagnosis the covering condition must be satisfied.

We examine the correspondence between abductive diagnosis and the notion of weak-causality diagnosis by an example.

Example 33. Reconsider the abductive diagnostic problem $\mathcal{A} = (\mathcal{C}, E)$ with the causal specification $\mathcal{C}$ from Section 2.1, and $E = \{f_1, f_2\}$. Assumption literals are added to the individual axioms in $\mathcal{R}$, yielding the causal specification $\mathcal{C}' = (\Delta', \Phi, \mathcal{R}')$, with $\mathcal{R}'$ equal to:

$$\begin{aligned} d_1 \wedge \alpha_1 &\rightarrow d_2 \\ d_1 \wedge \alpha_2 &\rightarrow f_1 \\ d_2 \wedge \alpha_3 &\rightarrow f_2 \\ d_2 \wedge d_3 \wedge \alpha_4 &\rightarrow f_3. \end{aligned}$$

The resulting evidence function is again equal to e as defined in Example 5.

Now, the set $H = \{d_1, \alpha_1, \alpha_2, \alpha_3\}$ is an abductive solution to $\mathcal{A}' = (\mathcal{C}', E)$, because the covering and consistency conditions are satisfied; the associated diagnosis is $D = \{d_1\}$. We also have that $\mathrm{WC}_{\Sigma,e_{|\{d_1\}}}(E) = \{d_1\}$.

If we restrict the notion of weak-causality diagnosis to monotonically increasing evidence functions, which is similar to restricting to standard logic in the theory of abductive diagnosis, the notion of diagnosis WC is diagnostically monotonic. This can be shown by noting that if $\mathrm{WC}_{\Sigma,e_{|H}}(E) = H$ and $H' \supset H$ then $\mathrm{WC}_{\Sigma,e_{|H'}}(E) = H'$, because if $e_{|H}(H) \supseteq E$ then $e_{|H'}(H') \supseteq E$, due to the fact that the evidence function e is monotonically increasing. Since only part of all observed findings may be accounted for by a subset of a set of defects H, where H accounts for all observed findings, the independence assumption, however, fails to hold.

Weak-causality diagnosis can be viewed as a much generalized version of set-covering diagnosis as defined in [29]; when an evidence function is assumed to be interaction-free, the two notions coincide.

Until now, weakly and strongly causal knowledge and their use in abductive diagnosis have been studied separately. Weak and strong causality diagnosis, however, can also be combined to obtain a notion of diagnosis that combines these two different interpretations of causal knowledge. Firstly, the evidence function e in a diagnostic specification is split up into two evidence functions:

$$\nu : \wp(\Delta) \rightarrow \wp(\Phi) \cup \{\bot\}$$

called the *strong evidence function*, and

$$\alpha : \wp(\Delta) \rightarrow \wp(\Phi) \cup \{\bot\}$$

called the *weak evidence function*, and the functions ν and α are defined such that

$$e(D) = \begin{cases} \nu(D) \cup \alpha(D) & \text{if } \nu(D), \alpha(D) \neq \bot, \\ \bot & \text{otherwise,} \end{cases}$$

for each $D \subseteq \Delta$. The set of abnormality axioms with incompleteness assumption literals is interpreted by a weak evidence function, again by discarding assumption literals; abnormality axioms without assumption literals are interpreted by a strong evidence function.

To capture the joint effect of strong and weak causality on diagnostic problem solving, the results of two separate diagnostic functions must be combined. However, diagnostic functions capturing abductive diagnosis using strongly causal knowledge or weakly causal knowledge each operate on parts of a diagnostic specification. To describe a diagnostic specification as consisting of a collection of diagnostic specifications, the notion of modularization appears to be convenient.

Definition 34 (*Modularization*). A *modularization* $\mathcal{M}_\Sigma$ of a diagnostic specification $\Sigma = (\Delta, \Phi, e)$ is a finite set of diagnostic specifications $\mathcal{M}_\Sigma = \{\Sigma_1, \ldots, \Sigma_n\}$, where $\Sigma_i = (\Delta, \Phi, e_i)$, $1 \leqslant i \leqslant n$, $n \geqslant 1$, such that for each $D \subseteq \Delta$:

$$e(D) = \begin{cases} \bigcup_{i=1}^{n} e_i(D) & \text{if } e_i(D) \neq \bot,\ 1 \leqslant i \leqslant n, \\ \bot & \text{otherwise.} \end{cases}$$

Modularization of a diagnostic specification is now employed to define the composition of two diagnostic functions.

Definition 35 (*Composition of diagnostic functions*). Let P, Q and R be three notions of diagnosis, and let $\mathcal{M}_\Sigma = \{\Sigma', \Sigma''\}$ be a modularization of the diagnostic specification Σ. Then, the diagnostic function $P_{\Sigma, e_{|H}}$ is called the *composition* of $Q_{\Sigma', e'_{|H}}$ and $R_{\Sigma'', e''_{|H}}$, denoted by

$$P_{\Sigma, e_{|H}} = Q_{\Sigma', e'_{|H}} \| R_{\Sigma'', e''_{|H}}$$

if it holds that

$$P_{\Sigma, e_{|H}}(E) = Q_{\Sigma', e'_{|H}}(E') \cup R_{\Sigma'', e''_{|H}}(E'')$$

for each set of observed findings $E \subseteq \Phi$, and each decomposition $E = E' \cup E''$ for which $Q_{\Sigma', e'_{|H}}(E')$, $R_{\Sigma'', e''_{|H}}(E'') \neq u$; otherwise $P_{\Sigma, e_{|H}}(E) = u$.

Observe that the sets E' and E'' resulting from a decomposition of the set of observed findings E are neither necessarily disjoint nor unique. Note also that the hypothesis H is the same for all diagnostic functions in a composition. This prerequisite ensures that possible dependencies among the respective evidence functions e' and e'' are dealt with adequately.

Using the translation scheme and the composition of diagnostic functions, the following notion of diagnosis fully captures the theory of abductive diagnosis. The resulting notion of diagnosis is called weak-and-strong causality diagnosis, abbreviated to WSC. Let

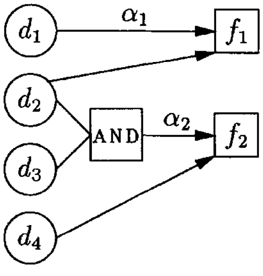


Fig. 10. Causal net corresponding to $\mathcal{C}$.

$\mathcal{M}_\Sigma = \{\Sigma_\nu, \Sigma_\alpha\}$ be a modularization of a diagnostic specification $\Sigma = (\Delta, \Phi, e)$, where $\Sigma_\nu = (\Delta, \Phi, \nu)$ and $\Sigma_\alpha = (\Delta, \Phi, \alpha)$. The notion of *weak-and-strong causality diagnosis*, denoted by WSC, is defined as follows:

$$\mathrm{WSC}_{\Sigma, e|_H} = \mathrm{SC}_{\Sigma_\nu, \nu|_H} \| \mathrm{WC}_{\Sigma_\alpha, \alpha|_H}$$

where SC is the notion of strong-causality diagnosis, and WC is the notion of weak-causality diagnosis.

Example 36. Consider the following abductive diagnostic problem $\mathcal{A} = (\mathcal{C}, E)$, with causal specification $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$, where $\mathcal{R}$ is equal to:

$$\begin{aligned} d_1 \wedge \alpha_1 &\rightarrow f_1 \\ d_2 \wedge d_3 \wedge \alpha_2 &\rightarrow f_2 \\ d_2 &\rightarrow f_1 \\ d_4 &\rightarrow f_2 \end{aligned}$$

$\Delta_P = \{d_1, d_2, d_3, d_4\}$, and $\Phi_P = E = \{f_1, f_2\}$. The causal specification $\mathcal{C} = (\Delta, \Phi, \mathcal{R})$ is graphically depicted in Fig. 10. The following modularization $\mathcal{M}_\Sigma = \{\Sigma_\nu, \Sigma_\alpha\}$ can be constructed: $\Sigma_\nu = (\Delta, \Phi, \nu)$, where the bottom-up partial specification $\tilde{\nu}$ of ν is defined as follows:

$$\tilde{\nu}(D) = \begin{cases} \{f_1\} & \text{if } D = \{d_2\}, \\ \{f_2\} & \text{if } D = \{d_4\}, \\ \emptyset & \text{if } D = \{d_i\},\ i = 1, 3,\ \text{or } D = \{\neg d_i\},\ i = 1, \ldots, 4. \end{cases}$$

Furthermore, $\Sigma_\alpha = (\Delta, \Phi, \alpha)$, where the bottom-up partial specification $\tilde{\alpha}$ is defined as follows:

$$\tilde{\alpha}(D) = \begin{cases} \{f_1\} & \text{if } D = \{d_1\}, \\ \{f_2\} & \text{if } D = \{d_2, d_3\}, \\ \emptyset & \text{if } D = \{d_i\},\ i = 2, 3, 4,\ \text{or } D = \{\neg d_i\},\ i = 1, \ldots, 4. \end{cases}$$

Since every observable finding in $e(D)$ is positive, only positive findings will be dealt with. An example of a diagnostic function comprising the notion of weak-and-strong causality diagnosis WSC is

$$\mathrm{WSC}_{\Sigma, e|_{\{d_1, d_2\}}} = \mathrm{WC}_{\Sigma_\alpha, \alpha|_{\{d_1, d_2\}}} \| \mathrm{SC}_{\Sigma_\nu, \nu|_{\{d_1, d_2\}}}.$$

Note that, for example, $\mathrm{WSC}_{\Sigma,e|\{d_1,d_2\}}(\emptyset) = u$, because $\mathrm{SC}_{\Sigma_\nu,\nu|\{d_1,d_2\}}(\emptyset) = u$, although $\mathrm{WC}_{\Sigma_\alpha,\alpha|\{d_1,d_2\}}(\emptyset) = \{d_1, d_2\}$. Observe also that a set of observed findings E may be decomposed among diagnostic functions of WC and SC in several ways. For example,

$$\begin{aligned}\mathrm{WSC}_{\Sigma,e|\{d_2,d_3,d_4\}}(\{f_1, f_2\}) &= \mathrm{WC}_{\Sigma_\alpha,\alpha|\{d_2,d_3,d_4\}}(\emptyset) \cup \mathrm{SC}_{\Sigma_\nu,\nu|\{d_2,d_3,d_4\}}(\{f_1, f_2\})\\ &= \mathrm{WC}_{\Sigma_\alpha,\alpha|\{d_2,d_3,d_4\}}(\{f_2\}) \cup \mathrm{SC}_{\Sigma_\nu,\nu|\{d_2,d_3,d_4\}}(\{f_1, f_2\})\\ &= \mathrm{WC}_{\Sigma_\alpha,\alpha|\{d_2,d_3,d_4\}}(\{f_1\}) \cup \mathrm{SC}_{\Sigma_\nu,\nu|\{d_2,d_3,d_4\}}(\{f_1, f_2\})\\ &= \mathrm{WC}_{\Sigma_\alpha,\alpha|\{d_2,d_3,d_4\}}(\{f_1, f_2\}) \cup\\ &\quad\ \mathrm{SC}_{\Sigma_\nu,\nu|\{d_2,d_3,d_4\}}(\{f_1, f_2\}).\end{aligned}$$

4.2. Consistency-based diagnosis

In consistency-based diagnosis, as proposed in [36] and [18] and introduced in Section 2.1, knowledge concerning structure and behaviour of a device is represented as a pair $\mathcal{S} = (\mathrm{SD}, \mathrm{COMPS})$, called a *system*; when observed findings OBS are included, we arrive at what is called an *observed system* $\mathrm{OS} = (\mathcal{S}, \mathrm{OBS})$, where

- SD denotes a finite set of formulae in first-order predicate logic, specifying normal structure and behaviour, called a *system description*, or sometimes also *normality axioms*;
- COMPS denotes a finite set of constants in first-order logic, denoting the *components* (elements) of the system;
- OBS 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. Adding knowledge of abnormal behaviour can be an effective means to reduce the number of alternative diagnoses produced [18].

A consistency-based diagnosis is defined as an assignment of either a positive literal Abnormal(c) or a negative literal ¬Abnormal(c) to each $c \in \mathrm{COMPS}$, i.e.,

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

where $C \subseteq \mathrm{COMPS}$, such that

$$\mathrm{SD} \cup \mathrm{OBS} \cup D \nvDash \bot$$

($\mathrm{SD} \cup \mathrm{OBS} \cup D$ is satisfiable); this condition is called the *consistency condition*. In the formalization by de Kleer et al., each literal $\mathrm{Abnormal}(c) \in D$ is interpreted as being defective; a literals $\neg\mathrm{Abnormal}(c) \in D$ indicates component c to be nondefective [18]. In the original theory by Reiter, the set C above is taken as a diagnosis, with the extra requirement that C is minimal with respect to set inclusion [36], but note that for each component $c \in C$, it holds that Abnormal(c) is *true*, i.e., c is defective. According to the definition of consistency-based diagnosis, taking $C = \mathrm{COMPS}$ leads to the trivial diagnosis that all components are defective (or the defective components are among the set of all components). This explains why Reiter incorporated in the original theory the requirement that the set C must be a minimal set with respect to set inclusion, fulfilling the consistency condition. However, later it was recognized 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.

Example 37. Consider the system $\mathcal{S} = (\mathrm{SD}, \mathrm{COMPS})$ from Section 2.1 (Fig. 2). Now, let $\mathrm{OBS} = \{I_1 = 1, I_2 = 0, I_3 = 1, O_1 = 1, O_2 = 0\}$, then

$$D = \{\neg\mathrm{Abnormal}(X), \mathrm{Abnormal}(A)\}$$

is a diagnosis, because

$$\mathrm{SD} \cup \mathrm{OBS} \cup D \nvDash \bot,$$

i.e., consistency has been regained by assuming the AND gate A to be faulty, whereas assuming both X and A to be normally functioning, i.e.,

$$D' = \bigl\{\neg\mathrm{Abnormal}(X), \neg\mathrm{Abnormal}(A)\bigr\}$$

yields an inconsistency ($\mathrm{SD} \cup \mathrm{OBS} \cup D' \vDash \bot$), indicating that D' is no diagnosis.

In [18] the notion of *partial diagnosis* is introduced, which is a satisfiable set D of Abnormal(c) and ¬Abnormal(c) assignments to part of all the components (the abnormality of the remaining components is thus assumed to be unknown), such that the consistency condition is fulfilled for every satisfiable superset of D. A *kernel diagnosis* is a partial diagnosis that is minimal with respect to set inclusion, and can be viewed as denoting a common diagnostic pattern.

This notion of diagnosis can be defined in terms of our framework. The resulting notion of *consistency-based diagnosis*, denoted by CB, is defined as follows:

$$\mathrm{CB}_{\Sigma,e_{|H}}(E) = \begin{cases} H & \text{if } \forall f \in E\colon\ f \in e_{|H}(H) \vee \neg f \notin e_{|H}(H), \\ u & \text{otherwise.} \end{cases}$$

A hypothesis H may also include observable findings as inputs to a system, in which case H is a set of assumptions concerning findings and defects.

Example 38. For the logic circuit in Fig. 2 we have that

$$\mathrm{CB}_{\Sigma,e'_{|\{\neg x,a\}}}\bigl(\{o_1, \neg o_2\}\bigr) = \{\neg x, a\},$$

where $\neg x$ means that the XOR gate X is normal and a means that the AND gate A is abnormal or faulty (cf. Example 6 for the evidence function e'). This result is analogous to the diagnosis in Example 37, obtained by the corresponding logical definition of consistency-based diagnosis.

Without further restrictions with regard to the evidence functions e, the notion of CB diagnosis is neither diagnostically monotonic nor is the independence assumption satisfied. However, the independence assumption is satisfied if CB diagnosis is restricted to diagnostic specifications that are monotonically increasing. As discussed in Section 2, evidence functions representing system descriptions are typically monotonically decreasing. If the notion of diagnosis CB is defined for such functions, the independence assumption fails to hold, as can be shown by a simple counterexample. The following useful proposition holds in case the evidence function is monotonically decreasing.

Proposition 39. *Let $\mathcal{P} = (\Sigma, E)$, $\Sigma = (\Delta, \Phi, e)$, be a diagnostic problem with monotonically decreasing evidence function e, and let $H \supseteq H'$, with $H, H' \subseteq \Delta$, then if $\mathrm{CB}_{\Sigma, e_{|H}}(E) = D$, then $\mathrm{CB}_{\Sigma, e_{|H'}}(E) = D'$ with $D' \subseteq D$.*

Proof. If $\mathrm{CB}_{\Sigma, e_{|H}}(E) = D$ then for each $f \in E$: (1) $f \in e_{|H}(H)$ or (2) $\neg f \notin e_{|H}(H)$. If condition (1) holds then $f \in e_{|H'}(H')$, because $e_{|H}(H) \subseteq e_{|H'}(H')$; for the same reason from condition (2) it follows that $\neg f \notin e_{|H'}(H')$. □

In terms of the approach by de Kleer et al. [18], from this proposition the existence of a partial diagnosis can be derived.

Corollary 40. *Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem, with monotonically decreasing evidence function e, then if $\mathrm{CB}_{\Sigma, e_{|H \cup \{d\}}}(E) = H \cup \{d\}$ and $\mathrm{CB}_{\Sigma, e_{|H \cup \{\neg d\}}}(E) = H \cup \{\neg d\}$, then also $\mathrm{CB}_{\Sigma, e_{|H}}(E) = H$.*

In [18], the notion of partial diagnosis is provided as a basic definition; it is not derived from the notion of diagnosis, as done above.

4.3. Comparison

It is informative to relate the notions of diagnosis introduced above to each other in terms of the restriction relation $\sqsubseteq$ (cf. Definition 28). It is easily seen that the following proposition holds.

Proposition 41. *Let* SC, WC *and* CB *be the notions of strong-causality, weak-causality and consistency-based diagnosis, respectively, then*

$$\mathrm{SC} \sqsubseteq \mathrm{WC} \sqsubseteq \mathrm{CB}.$$

Proof. Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem. Simply observe that if $\mathrm{WC}_{\Sigma, e_{|H}}(E) = H$, then $e_{|H}(H) = E$, therefore, $e_{|H}(H) \supseteq E$, and $\mathrm{WC}_{\Sigma, e_{|H}}(E) = H$. Furthermore, if $\mathrm{WC}_{\Sigma, e_{|H}}(E) = H$, then $e_{|H}(H) \supseteq E$, so for each $f \in E$: $f \in e_{|H}(H)$. Hence, $\mathrm{CB}_{\Sigma, e_{|H}}(E) = H$ holds. □

This reveals that consistency-based diagnosis is a very weak form of diagnosis, potentially producing many alternative diagnoses, a well known fact in the diagnosis community.

5. Refinement diagnosis

Although the diagnostic theories mentioned above differ in several respects, diagnostic problem solving can be viewed in all of them as a special instance of hypothetical reasoning

[31]. In solving a diagnostic problem, a hypothesis is first generated and next tested with respect to diagnostic knowledge and observed findings. If it passes the tests, it is accepted and called a diagnosis; when it fails to pass the tests, it is rejected.

This view of diagnosis is quite general, but it is still unnecessarily restrictive. Instead of simply rejecting a hypothesis that does not comply with all requirements, it seems more natural to adjust or refine it, when possible. Then, a diagnosis obtained after refinement of a hypothesis may be viewed as the best possible solution in a particular sense, given the domain knowledge, the set of observed findings and the hypothesis at hand. It therefore seems attractive to incorporate a principle of refinement into the basic definition of diagnosis, yielding notions of *refinement diagnosis*. The study of these notions of diagnosis demonstrates the flexibility of the framework of diagnosis defined in Sections 2 and 3.

There are various reasons why refinement diagnosis may be a more appropriate basis for diagnostic problem solving than the more rigorous notions of diagnosis mentioned above:

- Real-world knowledge bases are, almost without exception, incomplete, i.e., the modelled problem domain has not been fully described. For example, knowledge of certain interactions among defects may be missing.
- Real-world knowledge bases are not completely accurate, e.g., the meaning of the domain knowledge may not have been captured precisely.
- The findings that may be observed, and interpreted by a diagnostic system, are only part of what might have been collected without limitations, such as available time and money.
- Part of the observed findings may be unreliable, due to impediments to the observation process, such as limited available time.

In many domains, in particular medicine, it is usually better to arrive at a diagnosis that does not account for all observed findings, or that suggests findings that have not been observed, than to establish no diagnosis at all. It is sometimes said that such a diagnosis *underaccounts* or *overaccounts* for the set of observed findings.

The following question now arises: what can be taken as a basis for notions of diagnosis which incorporate certain principles of refinement? Obviously, there exists a wide range of possibilities. Which of the possible choices yields the most natural result depends, to a large extent, on the nature of the problem domain, which is partially expressed by the characteristics of the evidence functions e. Dependencies between a notion of diagnosis R, on the one hand, i.e., the interpretation of the set of observed findings given a specific knowledge base, and properties of a given evidence function e, on the other hand, are of importance in this respect.

Two classes of refinement diagnosis will be studied here. Firstly, the class of notions of refinement diagnosis, called *most general diagnosis*, is examined, where the least upper bound of accepted hypotheses (with respect to set inclusion) is taken as a diagnostic solution. Secondly, the class of notions of refinement diagnosis, called *most specific diagnosis*, based on taking the greatest lower bound of accepted hypotheses is studied. In most general diagnosis, the smallest set of defects that includes every accepted subhypothesis is considered most plausible; in contrast, in most specific diagnosis, the largest set of defects that is included in every accepted subhypothesis is considered most plausible.

5.1. Most general diagnosis

Notions of most general diagnosis capture the idea that if a specific diagnostic hypothesis is not accepted, then the 'nearest' subhypothesis should be taken instead. The least upper bound with respect to set inclusion of the set of accepted subhypotheses is an example of such a 'nearest' subhypothesis.

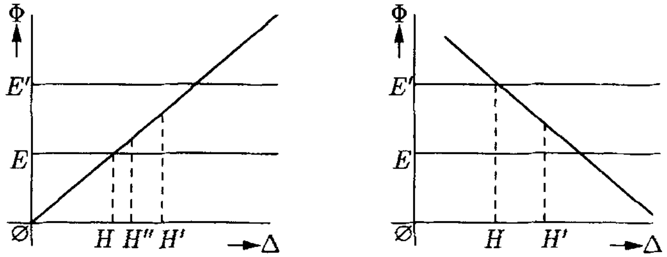
The notion of *most general subset diagnosis*, denoted by GS, is defined as follows:

$$\mathrm{GS}_{\Sigma,e_{|H}}(E) = \begin{cases} \bigcup\limits_{\substack{H' \subseteq H \\ e_{|H}(H') \subseteq E}} H' & \text{if } H \text{ is consistent, and } \exists H' \subseteq H\colon e_{|H}(H') \subseteq E, \\ u & \text{otherwise.} \end{cases}$$

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 42. Reconsider the causal specification C in Fig. 1 and the corresponding evidence function e in Example 5, with $E = \{f_1, f_2\}$ (fever and sore throat), we have that $\mathrm{GS}_{\Sigma,e_{|\{d_1,d_2\}}}(E) = \{d_1, d_2\}$, which is also an abductive diagnosis, because $\mathrm{SC}_{\Sigma,e_{|\{d_1,d_2\}}}(E) = \{d_1, d_2\}$. However, it holds that $\mathrm{GS}_{\Sigma,e_{|\{d_1,d_2\}}}(\{f_2\}) = \{d_2\}$, where $\mathrm{SC}_{\Sigma,e_{|\{d_1,d_2\}}}(\{f_2\}) = u$. Hence, $e(\{d_1, d_2\})$ predicts a finding that cannot be accounted for, causing the defect d_1 to be ignored. This may be a suitable approach to domains in which neglecting a particular defect may be dangerous.

In Fig. 11, the relationship between diagnostic hypothesis H, the set of observed findings E and the resulting diagnosis $\mathrm{GS}_{\Sigma,e_{|H}}(E)$ is summarized by schematically depicting these sets as if they were real numbers and by taking set inclusion as the $\leqslant$ total order on the real numbers. If most general subset diagnosis is applied to a monotonically decreasing evidence function, the resulting diagnosis is either undefined or equal to



$\mathrm{GS}_{\Sigma,e_{|H}}(E) = H$
$\mathrm{GS}_{\Sigma,e_{|H'}}(E) = H''$
$\mathrm{GS}_{\Sigma,e_{|H'}}(E') = H'$

(a)

$\mathrm{GS}_{\Sigma,e_{|H}}(E') = H$
$\mathrm{GS}_{\Sigma,e_{|H'}}(E') = H'$
$\mathrm{GS}_{\Sigma,e_{|H'}}(E) = u$

(b)

Fig. 11. Monotonically increasing (a) and decreasing (b) evidence functions.

the given hypothesis H. This contrasts with GS applied to a monotonically increasing evidence function, which may also yield subsets of the hypothesis as a diagnosis. If such an evidence function is assumed to represent empirical associations, the notion of most general diagnosis may be taken as the formalization of heuristic classification. $\mathrm{GS}_{\Sigma,e_{|H'}}(E) = H''$ in Fig. 11(a) is intended to illustrate that $e(H'')$ may even be a superset of E. If the evidence function e is nonmonotonic, then the relationships between E and $e_{|H}(H')$ are investigated as before, but again, certain interactions between defects may be ignored.

The independence assumption is satisfied for GS if GS is restricted to diagnostic specifications with a monotonically increasing evidence function, which can be defined by a bottom-up partial specification.

Proposition 43. *The independence assumption holds for the notion of diagnosis* GS, *when applied to diagnostic specifications with monotonically increasing evidence functions, described by a bottom-up partial specification.*

Proof. Let $\mathcal{P} = (\Sigma, E)$ be a diagnostic problem with monotonically increasing evidence function e. Let $V \subseteq H$ be a subset of the hypothesis $H \subseteq \Delta$. The powerset $\wp(H)$ is partitioned into the set of sets P for which it holds that for each $U \in P$: $U \subseteq V$, and the set of sets P' for which it holds that for each $U \in P'$: $U \not\subseteq V$. Then, according to basic set theory, it holds that:

$$\mathrm{GS}_{\Sigma,e_{|H}}(E) = \bigcup_{\substack{H' \in P \\ e_{|V}(H') \subseteq E}} H' \cup \bigcup_{\substack{H' \in P' \\ e_{|H}(H') \subseteq E}} H'.$$

The first component of this union can also be written as $\mathrm{GS}_{\Sigma,e_{|V}}(E)$. Since e is monotonically increasing, the sets $H' \in P'$ may be changed to $H'' = H' \backslash V$, because if $e(H') \subseteq E$, then $e(H'') \subseteq E$, and because $H' \cap V \subseteq V$, the set $H' \cap V$ is considered in the diagnosis $\mathrm{GS}_{\Sigma,e_{|V}}(E)$. Hence,

$$\mathrm{GS}_{\Sigma,e_{|H}}(E) = \mathrm{GS}_{\Sigma,e_{|V}}(E) \cup \mathrm{GS}_{\Sigma,e_{|H\backslash V}}(E).$$

Since the set V has been selected arbitrarily, GS satisfies the independence assumption. □

However, if the evidence function e is not monotonically increasing, then the independence assumption is not satisfied. Hence, the independence assumption fails to hold in general for most general subset diagnosis, as can be shown the a counterexample. However, most general subset diagnosis is diagnostically monotonic, as proven in the following proposition.

Proposition 44. *The notion of most general subset diagnosis* GS *is diagnostically monotonic.*

Proof. If $H \subseteq H'$, then $\mathrm{GS}_{\Sigma,e_{|H}}(E) \subseteq \mathrm{GS}_{\Sigma,e_{|H'}}(E)$ given that $\mathrm{GS}_{\Sigma,e_{|H}}(E), \mathrm{GS}_{\Sigma,e_{|H'}}(E) \neq u$, because if $e_{|H}(H'') \subseteq E$, $H'' \subseteq H$, then $e_{|H'}(H'') \subseteq E$. □

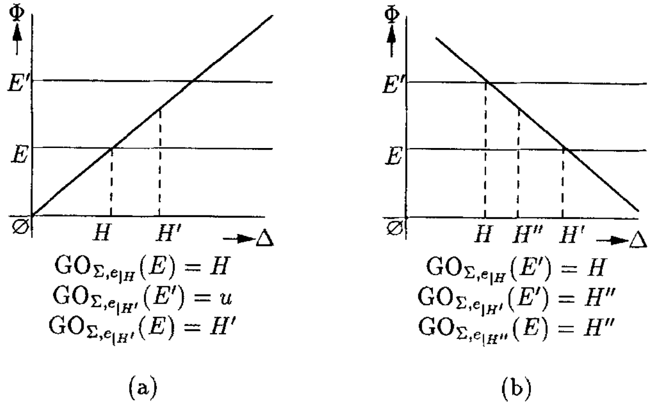


Fig. 12. Monotonically increasing (a) and decreasing (b) evidence functions.

Where most general subset diagnosis can be viewed as a more flexible version of strong-causality diagnosis SC, which for certain evidence functions is as little restrictive as consistency-based diagnosis, a similar, flexible notion of diagnosis can be designed for weak-causality diagnosis. This suggests replacing the subset relation in most general subset diagnosis by the superset relation, yielding the notion of most general superset diagnosis GO (the letter 'O' stands for 'cOntains').

The notion of *most general superset diagnosis*, denoted by GO, is defined as follows:

$$\mathrm{GO}_{\Sigma,e_{|H}}(E) = \begin{cases} \displaystyle\bigcup_{\substack{H' \subseteq H \\ e_{|H}(H') \supseteq E}} H' & \text{if } H \text{ is consistent, and } \exists H' \subseteq H\colon\ e_{|H}(H') \supseteq E, \\ u & \text{otherwise.} \end{cases}$$

Most general superset diagnosis has much in common with weak-causality diagnosis WC defined in the previous section. If the notion of most general superset diagnosis is applied to evidence functions that are monotonically decreasing, or nonmonotonic, for the resulting diagnosis $\mathrm{GO}_{\Sigma,e_{|H}}(E) = H'$ it may even hold that $e(H') \subset E$, although for each of the diagnostic hypotheses $H'' \subseteq H$ that contribute to the diagnosis it holds that $e_{|H}(H'') \supseteq E$. Hence, the situation is the reverse of that for most general subset diagnosis discussed above, as might be expected from their respective definitions. In Fig. 12, the various possibilities are schematically depicted. The independence assumption is not generally satisfied for most general superset diagnosis, but most general superset diagnosis is diagnostically monotonic. Both results follow from straightforward modification of Proposition 44.

As is true for weak-causality diagnosis WC, most general superset diagnosis restricted to monotonically increasing evidence functions is very unrestrictive, which is revealed by the fact that $\mathrm{GO}_{\Sigma,e_{|H}}(\emptyset) = H$ if $e(H) \neq \bot$, meaning that all defects constituting the hypothesis may have occurred, even if no findings have been observed. Note that the same diagnosis would have been produced by weak-causality diagnosis WC in this case. By adopting some criterion of parsimony to only select plausible diagnoses (cf. [41]), such as minimality according to set inclusion, the unrestrictiveness is alleviated; the empty diagnosis $\emptyset$ would then be produced.

An alternative to the definition of subset diagnosis is to consider all sets of defects D that have at least one finding f in common with the findings E observed. This leads to the following definition of the notion of *most general intersection diagnosis*, denoted by GI:

$$\mathrm{GI}_{\Sigma,e_{|H}}(E)=\begin{cases}\displaystyle\bigcup_{\substack{H'\subseteq H\\ (E=\emptyset\vee e_{|H}(H')=\emptyset\,\vee\\ e_{|H}(H')\cap E\neq\emptyset)}} H' & \text{if } H \text{ is consistent, and } (E=\emptyset\\ & \text{or } \exists H'\subseteq H\text{: } e_{|H}(H')=\emptyset \text{ or}\\ & e_{|H}(H')\cap E\neq\emptyset),\\ u & \text{otherwise.}\end{cases}$$

If the sets of observed and observable findings are nonempty, intersection diagnosis with respect to H stands for the least upper bound of subsets of defects of $H \subseteq \Delta$, where for each subset of defects H' admitted to the most general intersection diagnosis $\mathrm{GI}_{\Sigma,e_{|H}}(E)$, the associated set of observable findings $e_{|H}(H')$ is empty or has at least one finding in common with the set of observed findings E.

The independence assumption is not satisfied for most general intersection diagnosis, which is even true if GI is restricted to interaction-free evidence functions. The reason is that if $e_{|H}(H')\cap E\neq\emptyset$, then it need not be true that for all $d\in H'$: $e_{|\{d\}}(d)\cap E\neq\emptyset$. Only if $e(D)=e(D')$, for each consistent $D, D'\subseteq\Delta$ (every set of defects has the same set of associated findings) would the independence assumption hold. However, if e is interaction free, the notion of most general intersection diagnosis restricted to such interaction-free evidence functions is diagnostically monotonic.

The advantage of most general intersection diagnosis over most general subset and superset diagnosis is that only defects that have at least one associated observable finding that has actually been observed, are included in a diagnosis. This will be an acceptable assumption in a domain where not all findings associated with a set of defects need be observed and not all observed findings need be accounted for. In representing a domain, it may be required to restrict to those observable findings that are in some way 'typical' for the defects.

Most general intersection diagnosis can be viewed as a refinement version of a mixture of the notions of weak-causality and strong-causality diagnosis.

5.2. Comparison

Most general subset, superset and intersection diagnosis are three refinement approaches to diagnosis. The restriction relationships between these notions of diagnosis are shown in Fig. 13. For most general subset diagnosis, all findings associated with a set of defects must be observed if the set of defects is to be included as part of a diagnosis. Most general superset diagnosis focusses on common findings of defects. For most general intersection

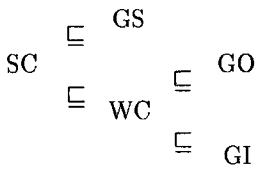


Fig. 13. Restriction taxonomy of notions of diagnosis.

diagnosis, at least one finding associated with a defect must be observed if the defect is to be included as part of a diagnosis. Notions of diagnosis can also be classified in terms of the subdiagnostic relation $\trianglelefteq$ (cf. Definition 29). The three notions of diagnosis discussed above stand in a subdiagnostic relation to each other:

$$\begin{aligned} &\mathrm{GS} \trianglelefteq \mathrm{GI} \\ &\mathrm{GO} \trianglelefteq \mathrm{GI}. \end{aligned}$$

This follows from the fact that if a set of observed findings is included in the set of observable findings associated with a set of defects, or vice versa, the intersection of the set of observed findings and observable findings is nonempty, given that neither the set of observed findings E, nor the set of observable findings $e_{|H}(H')$, is empty. For the empty cases, the most general intersection diagnosis is always equal to the largest result with respect to set inclusion of GO and GS. Hence, a most general intersection diagnosis will always contain at least as many elements as the corresponding most general subset and superset diagnosis.

5.3. *Most specific diagnosis*

Rather than taking the least upper bound of a set of accepted subhypotheses of a given hypothesis, taking the greatest lower bound provides another approach to refinement diagnosis. We shall refer to notions of diagnosis based on taking the greatest lower bound as notions of *most specific diagnosis*. Where the concept of most general diagnosis formalizes notions of diagnosis that yield diagnoses that include *every* accepted subhypothesis, most specific diagnosis formalizes notions of diagnosis that yield diagnoses that are *common* to every accepted subhypothesis. In general it holds for a notion of most specific diagnosis S that if $S_{\Sigma,e_{|H}}(E) = \emptyset$ and $S_{\Sigma,e_{|H'}}(E) = H''$, then, by definition, $S_{\Sigma,e_{|H\cup H'}}(E) = \emptyset$. Hence, notions of most specific diagnosis are very restrictive, and neither the independence assumption nor diagnostic monotonicity holds.

As with the notion of most general subset diagnosis, in the notion of most specific subset diagnosis, subhypotheses are admitted to a diagnosis if their associated sets of findings are included in the set of observed findings of a diagnostic problem. However, of these accepted subhypotheses, only the defects the subhypotheses have in common constitute a diagnosis. Hence, the notion of *most specific subset diagnosis*, denoted by SS, is defined as follows:

$$\mathrm{SS}_{\Sigma,e_{|H}}(E) = \begin{cases} \bigcap\limits_{\substack{H' \subseteq H \\ e_{|H}(H') \subseteq E}} H' & \text{if } H \text{ is consistent, and } \exists H' \subseteq H\colon e_{|H}(H') \subseteq E, \\ u & \text{otherwise.} \end{cases}$$

This notion of diagnosis is extremely restrictive. For example, if an evidence function is interaction free, then the most specific subset diagnosis will almost always (with the exception when only one subhypothesis is accepted) be equal to the empty set.

If the evidence function is monotonically decreasing, then most specific subset diagnosis tries to construct the smallest diagnosis possible. It may be viewed as a flexible form of kernel, consistency-based diagnosis in the sense of [18]. The reason for the

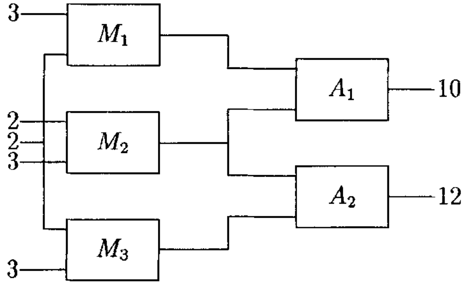


Fig. 14. Multiplier–adder circuit.

similarity between kernel diagnosis in consistency-based diagnosis and most specific subset diagnosis is that any hypothesis H' for which $e_{|H}(H') \subseteq E$ is also consistent with E.

The correspondence between kernel diagnosis and most specific subset diagnosis is illustrated by an example taken from [18].

Example 45. Consider Fig. 14, which depicts an electronic circuit with three multipliers, referred to as M_1, M_2 and M_3, and two adders, denoted by A_1 and A_2. When a multiplier M_i is behaving normally, it produces as output the product of its two inputs; similarly, a normally behaving adder A_j produces as output the sum of its two inputs. Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification representing the circuit. The fact that some multiplier M_i is defective, is denoted by m_i; if it is nondefective, this is indicated by $\neg m_i$. A similar notational convention is adopted with regard to the two adders. It is convenient to assume that the input to the circuit is fixed (as assumed in [15,18]), as indicated in Fig. 14. The normal output of the circuit, $O_1 = 12$ and $O_2 = 12$, is denoted by o_1 and o_2; abnormal output is denoted by $\neg o_j$, $j = 1, 2$. The set of observed findings E is in equal to $e = \{\neg o_1, o_2\}$, i.e., $O_1 = 10 \neq 12$ and $O_2 = 12$.

The following values of the evidence function e are among those that correspond to the circuit's normal behaviour:

$$\begin{aligned}
e(\{\neg m_1, \neg m_2, \neg m_3, \neg a_1, \neg a_2\}) &= \{o_1, o_2\} \\
e(\{\neg m_1, \neg m_2, \neg m_3, a_1, \neg a_2\}) &= \{o_2\} \\
e(\{\neg m_1, \neg m_2, \neg m_3, a_1, a_2\}) &= \emptyset \\
e(\{\neg m_1, \neg m_2, \neg m_3, a_1\}) &= \{o_2\} \\
e(\{a_1\}) &= \{o_2\} \\
e(\{\neg m_1, \neg m_2, \neg m_3\}) &= \{o_1, o_2\} \\
&\vdots \\
e(\emptyset) &= \{o_1, o_2\}.
\end{aligned}$$

($e(\emptyset)$ denotes that it is unknown whether defects are present or absent.) The most specific subset diagnosis with respect to the hypothesis $H = \{a_1\}$ is equal to

$$\mathrm{SS}_{\Sigma, e_{|\{a_1\}}}(\{\neg o_1, o_2\}) = \{a_1\}$$

which is indeed a kernel diagnosis for the diagnostic problem $\mathcal{P} = (\Sigma, E)$ using consistency-based diagnosis. Note that

$$\mathrm{SS}_{\Sigma,e_{|H}}\big(\{\neg o_1, o_2\}\big) = \{a_1\}$$

if $a_1 \in H$, when for the other kernel diagnoses it holds $\{m_1\}$, $\{m_2, m_3\}$, $\{m_2, a_2\} \not\subseteq H$, for example,

$$H = \{\neg m_1, \neg m_2, \neg m_3, a_1, \neg a_2\}.$$

As discussed above, most general superset diagnosis will often yield a diagnosis that contains too many defect elements, in particular when an evidence function is monotonically increasing. Most specific superset diagnosis is a more restrictive, and possibly more suitable, notion of diagnosis than most general superset diagnosis.

The notion of *most specific superset diagnosis*, denoted by SO, is defined as follows:

$$\mathrm{SO}_{\Sigma,e_{|H}}(E) = \begin{cases} \displaystyle\bigcap_{\substack{H' \subseteq H \\ e_{|H}(H') \supseteq E}} H' & \text{if } H \text{ is consistent, and } \exists H' \subseteq H\colon\ e_{|H}(H') \supseteq E, \\ u & \text{otherwise.} \end{cases}$$

If the evidence function to which most specific superset diagnosis is applied, is monotonically increasing, the result may be intuitively attractive. The basic idea of most specific superset diagnosis is that the observed findings that are common to the accepted subhypotheses are due to the common defects of accepted subhypotheses.

Example 46. Reconsider Fig. 1. For $E = \{f_2, f_3\}$ (i.e., the patient has sore throat and dyspnoea), the most specific superset diagnosis is equal to

$$\mathrm{SO}_{\Sigma,e_{|\{d_1,d_2,d_3\}}}(E) = \{d_3\}$$

because, it holds that $e_{|H}(\{d_1, d_3\}) \supseteq E$, $e_{|H}(\{d_2, d_3\}) \supseteq E$ and $e_{|H}(\{d_1, d_2, d_3\}) \supseteq E$, where $H = \{d_1, d_2, d_3\}$. All other subsets of H have associated sets of findings that are no supersets of E. The defect d_3 stands for asthma. While both d_1 and d_2 participate in subhypotheses that also account for E, only the defect d_3 occurs in all accepted subhypotheses, i.e., it turns out to be essential. It seems therefore intuitively right to accept d_3 as the most plausible diagnosis.

As the example above indicates, a most specific superset diagnosis need not account for all observed findings on the basis of the given evidence function. If an evidence function is interaction free, then most specific superset diagnosis is likely to produce a singleton set diagnosis for a given hypothesis that is very plausible if the associated sets of observed findings $e(\{d\})$ are mutually disjoint.

As discussed above, the notion of most general intersection diagnosis is very unrestrictive. All defects that, either individually or in combination with other defects, have findings in common with the set of observed findings, are included in a diagnosis. The notion of *most specific intersection diagnosis*, denoted by SI, is much more restrictive than most general intersection diagnosis; it is defined as follows:

$$\mathrm{SI}_{\Sigma,e_{|H}}(E) = \begin{cases} \displaystyle\bigcap_{\substack{H' \subseteq H \\ (E=\emptyset \vee e_{|H}(H')=\emptyset \vee \\ e_{|H}(H') \cap E \neq \emptyset)}} H' & \text{if } H \text{ is consistent, and } (E=\emptyset \text{ or} \\ & \exists H' \subseteq H\colon e_{|H}(H')=\emptyset \text{ or} \\ & e_{|H}(H') \cap E \neq \emptyset), \\ u & \text{otherwise.} \end{cases}$$

If the evidence function e is monotonically increasing, the resulting diagnosis will be equal to the empty set if the function values $e(\{d\})$ have many observable findings in common.

5.4. Comparison

Although the notions of most specific diagnosis are very restrictive, they do not stand in a simple restriction relation to the other notions of diagnosis. However, it is easy to see that

$$\mathrm{SS}_{\Sigma,e_{|H}}(E) \subseteq \mathrm{GS}_{\Sigma,e_{|H}}(E)$$

holds for each consistent $H \subseteq \Delta$. Similar set inclusion relations hold for the other notions of diagnosis. We state without proof that:

$$\begin{aligned} \mathrm{SS} &\trianglelefteq \mathrm{GS} \\ \mathrm{SO} &\trianglelefteq \mathrm{GO} \\ \mathrm{SI} &\trianglelefteq \mathrm{GI}. \end{aligned}$$

6. Discussion

In this paper, a theory of diagnosis has been developed which considers a diagnosis yielded by a diagnostic problem solver as an established relationship between interpreted domain knowledge and a hypothesis. The resulting framework is suitable to express *static* aspects of diagnosis, i.e., diagnosis without taking problem-solving strategies into account. It is inspired by the work on abductive diagnosis by Reggia et al. [29] and Bylander et al. [5], but offers a significant extension to that work. In fact, as has been shown above, these theories of abductive diagnosis amount to particular choices in our theory of diagnosis.

The framework of diagnosis proposed in this paper supports two different views. On the one hand, given some intuitively appealing interpretation of knowledge, expressed by an evidence function, a notion of diagnosis can be designed, or selected, that adheres to that meaning as closely as possible. On the other hand, applying a particular notion of diagnosis to solve a diagnostic problem implies that a particular (diagnostic) meaning is given to the associated evidence function. It was shown that well known notions of diagnosis from the literature are expressible in terms of the framework, and that it is suitable as a tool for the analysis and comparison of notions of diagnosis. Furthermore, several new notions of diagnosis have been proposed that are less rigorous in dealing with observed findings and diagnostic knowledge than common notions of diagnosis, which give up too soon, e.g., when a single element among the set of observed findings cannot be accounted for. However, these are certainly not the only notions of diagnosis that may be useful in certain domains.

The literature on diagnosis contains a number of other approaches to diagnosis. In particular, logic has been a popular language for the analysis of diagnosis, yielding a number of different logical notions of diagnosis, like abductive and consistency-based diagnosis [12,18,32,36,42]. These logical notions of diagnosis have usually been designed in close connection with specific domain models, such as causal models or models of structure and behaviour, and, hence, can be applied in a natural way to deal with specific diagnostic problems only. Although several researchers have demonstrated their theory of diagnosis to be more general than orginally thought [12,18,32,36], there remains a close link between a specific theory and problem type. In contrast, in our framework, there is no intimate connection between the theory and any of the existing conceptual models of diagnosis. In fact, the meaning of a knowledge base, described by means of an evidence function e, is completely separated from its diagnostic use. Of course, it is usually desirable to define notions of diagnosis that closely mirror the meaning of a knowledge base. Furthermore, where in other frameworks, modelled behaviour has to satisfy certain constraints, like monotonicity due to the monotonic nature of standard logical entailment, there are no such prerequisites in our framework, and many types of subtle interaction can be expressed.

We have focussed on qualitative methods for diagnostic problem solving, but in a considerable number of papers, diagnostic problem solving is essentially viewed as a form of reasoning with uncertainty, using specific quantitative measures of uncertainty. A typical example of such work is the usage of probabilistic networks, also called Bayesian belief networks, for diagnostic problem solving [1,20,28,39]. However, by a straightforward extension, the framework proposed in this paper can also cover such probabilistic diagnostic systems: Charniak and Shimony [7], and Santos and Santos [37,38], have shown that set-covering theory can be moved in a probabilistic direction, by the concept of *cost-based abduction*. This amounts to associating a prior cost function with sets of defects and findings, and updating cost information during abductive reasoning. Then, to any diagnosis produced, a cost will be attached. The cost of a diagnosis may be anything, varying from financial costs to some subjective feeling of importance expressed by numbers. However, Charniak and Shimony 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 using probabilistic networks [7].

A limitation of the framework presented here is that, as a tool for the semantical analysis of diagnosis, the framework is rather extensional in nature. This is in contrast with the more intensional nature of logic-based techniques for the analysis of diagnosis, as commonly used in consistency-based and abductive diagnosis, which allow for the separate specification of knowledge of structure and function, and for the easy composition of a knowledge base, just by putting parts together. Despite this limitation, it is the extensional nature of the formalism that forces one to think explicitly about interactions among defects and findings, and much insight can be gained in this way.

Acknowledgements

I would like to thank Ameen Abu-Hanna, Frank van Harmelen, John Josephson and Jan Treur for their many helpful comments on earlier drafts of this paper.

References

[1] S. Andreassen, M. Woldbye, B. Falck, S.K. Andersen, MUNIN—a causal probabilistic network for interpretation of electromyographic findings, in: Proceedings 10th International Joint Conference on Artificial Intelligence (IJCAI-87), Milan, Italy, 1987, pp. 366–372.

[2] A. Beschta, O. Dressler, H. Freitag, M. Montag, P. Struss, DPNet—a second generation expert system for localizing faults in power transmission networks, in: Proceedings International Conference on Fault Diagnosis (Tooldiag-93), Toulouse, France, 1993, pp. 1019–1027.

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

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

[5] T. Bylander. D. Allemang, M.C. Tanner, J.R. Josephson, The computational complexity of abduction, in: R.J. Brachman, H.J. Levesque, R. Reiter (Eds.), Knowledge Representation, The MIT Press, Cambridge, MA, 1992, pp. 25–60.

[6] B. Chandrasekaran, Generic tasks as building blocks for knowledge-based systems: the diagnosis and routine design examples, The Knowledge Engineering Review 3 (3) 183–210.

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

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

[9] L. Console, D. Theseider Dupré, P. Torasso, A theory of diagnosis for incomplete causal models, in: Proceedings 11th International Joint Conference on Artificial Intelligence, Detroit, MI, 1989, pp. 1311–1317.

[10] L. Console, P. Torasso, Integrating models of correct behaviour into abductive diagnosis, in: Proceedings ECAI-90, Stockholm, Sweden, 1990, pp. 160–166.

[11] L. Console, D. Theseider Dupré, P. Torasso, On the relationship between abduction and deduction, J. Logic Comput. 1 (5) (1991) 661–690.

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

[13] P.T. Cox, T. Pietrzykowski, General diagnosis by abductive inference, in: Proceedings IEEE Symposium on Logic Programming, 1987, pp. 183–189.

[14] P. Dague, Model-based diagnosis of analog electronic circuits, Ann. Math. Artificial Intelligence 11 (1994) 439–492.

[15] R. Davis, W. Hamscher, Model-based reasoning: troubleshooting, in: H.E. Shrobe (Ed.), Exploring Artificial Intelligence: Survey Talks from the National Conference on Artificial Intelligence, Morgan Kaufmann, San Mateo, CA, 1988, pp. 297–346.

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

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

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

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

[20] D.E. Heckerman, B.N. Nathwani, Towards normative expert systems: Part II—Probability-based representations for efficient knowledge acquisition and inference, Methods of Information in Medicine 31 (1992) 106–116.

[21] J.R. Josephson, S.G. Josephson, Abductive Inference: Computation, Philosophy, Technology, Cambridge University Press, Cambridge, 1994.

[22] K. Konolige, Using default and causal reasoning in diagnosis, Ann. Math. Artificial Intelligence 11 (1994) 97–135.

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

[24] P.J.F. Lucas, Logic engineering in medicine, The Knowledge Engineering Review 10 (2) (1995) 153–179.

[25] P.J.F. Lucas, Structures in diagnosis: from theory to medical application, Ph.D. Thesis, Free University of Amsterdam, Amsterdam, 1996.

[26] P.J.F. Lucas, Symbolic diagnosis and its formalisation, The Knowledge Engineering Review 12 (2) (1997) 109–146.

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

[28] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, San Mateo, CA, 1988.

[29] Y. Peng, J.A. Reggia, Abductive Inference Models for Diagnostic Problem Solving, Springer, New York, 1990.

[30] D. Poole, R. Goebel, R. Aleliunas, Theorist: a logical reasoning system for defaults and diagnosis, in: N. Cercone, G. McCalla (Eds.), The Knowledge Frontier, Springer, Berlin, 1987.

[31] D. Poole, A methodology for using a default and abductive reasoning system, Internat. J. Intelligent Systems 5 (5) (1990) 521–548.

[32] D. Poole, Representing diagnosis knowledge, Ann. Math. Artificial Intelligence 11 (1994) 33–50.

[33] C. Preist, K. Eshghi, B. Bertolino, Consistency-based and abductive diagnosis as generalized stable models, Ann. Math. Artificial Intelligence 11 (1994) 51–74.

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

[35] J.A. Reggia, D.S. Nau, Y. Wang, Diagnostic expert systems based on a set-covering model, Internat. J. Man-Machine Studies 19 (1983) 437–460.

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

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

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

[39] S.L. Lauritzen, D.J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, J. Roy. Statist. Soc. Ser. B 50 (2) (1987) 157–224.

[40] P. Torasso, L. Console, Diagnostic Problem Solving, North Oxford, London, 1989.

[41] S. Tuhrim, J. Reggia, S. Goodall, An experimental study of criteria for hypothesis plausibility, J. Experiment. Theoret. Artificial Intelligence 3 (1991) 129–144.

[42] A. ten Teije, F. van Harmelen, An extended spectrum of logical definitions for diagnostic systems, in: G.M. Provan (Ed.), DX-94, 5th International Workshop on Principles of Diagnosis, 1994, pp. 334–342.

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

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