An extended spectrum of logical definitions for diagnostic systems

Annette ten Teije, Frank van Harmelen
SWI
University of Amsterdam
{annette,frankh}@swi.psy.uva.nl

Abstract

The goal of this work is to develop a single uniform theory, which enables us to describe many different diagnostic systems. We will give a general definition of diagnostic systems. Our claim is that a large number of very different diagnostic systems can be described by this definition by choosing the right values for six parameters in this definition. Our work is an extension of the spectrum of logical definitions of Console and Torasso.

1 INTRODUCTION

Diagnosis is a complex task. For studying the diagnostic task it is useful to have a framework for describing diagnostic systems. A framework can help us in several ways: it makes classifying, comparing, choosing and reasoning about diagnostic systems easier because the framework enables us to express several diagnostic systems in the same manner. We will give a general definition of diagnostic systems. Our claim is that a large number of very different diagnostic systems can be described choosing the right values of six parameters in this definition. Although the framework is developed with a particular application in mind, namely for studying the diagnostic task, this paper is only about the framework itself and not about its use.

The question is what kind of framework do we develop: which aspects of diagnostic systems do we describe, in other words what are the necessary aspects for classifying, comparing, choosing, reasoning about diagnostic systems. We focus on the competence of diagnostic systems. This means that the input and output relation is the subject of study, i.e. aspects which depend on the functional specification. So we exclude aspects which depend on the algorithm or implementation of the diagnostic system.

The structure of this paper: Section 2 motivates why we take the spectrum of logical definitions for diagnostic systems of [Console & Torasso, 1991] as the starting point for our framework. Section 3 describes the notion of diagnosis and how we capture this notion in our framework. Section 4 gives two examples of diagnostic systems which are described in our extended framework. Section 5 emphasizes the difference between the original and our extended framework. Section 6 mentions several other proposed frameworks for diagnostic systems. In this section we explain why these frameworks are less appropriate starting points then the one of Console and Torasso. The last section is a summary and a description of our future work.

2 THE SPECTRUM OF CONSOLE AND TORASSO

In this section we explain why we take the framework of Console and Torasso as our starting point. There are several papers ([Leitch *et al.*, 1993, Böttcher *et al.*, 1991, Benjamins, 1993]) about a diagnostic theory. These papers try to give an overview of the various diagnostic systems by formulating these systems in a single vocabulary and developing a diagnostic theory in this vocabulary. The nature of these diagnostic theories differ, not only in language but also more fundamentally, in the aspects they describe (see the discussion about related work in section 6).

The spectrum of Console and Torasso is a framework for describing the relation between input and output of a diagnostic system. A diagnostic system is described by the set of observations Ψ^+ that has to be covered and a set of observations Ψ^- (which is computed from the observations and the observables) that may not contradict the explanation. The solution of a diagnostic problem are the explanations that have a minimal set of fault causes. The only parameter of the framework is Ψ^+. In other words, if there are two diagnostic systems which define Ψ^+ differently,

then these systems have different spectrum descriptions. For example, "pure abduction systems" a la Poole [Poole, 1989] take for Ψ^+ all the observations, and "pure consistency based systems" a la De Kleer [deKleer & Williams, 1987], [Reiter, 1987] take for Ψ^+ the empty set. These are also the two extremes of the spectrum.

We take this work of Console and Torasso as our starting point for several reasons. First it is a precise definition of diagnosis. A precise description is important for our application (reasoning about diagnostic systems) because automating such reasoning requires a formal definition. Second, the spectrum is appropriate for both streams in diagnosis, abduction based and consistency based, and not for only one of these two. The third reason is that the spectrum is a competence theory, namely a description of the input/output relation for diagnostic systems, as required for classifying, comparing, choosing and reasoning about diagnostic systems.

Console and Torasso capture a diagnostic system in a framework with only one parameter, namely the observations that have to be covered (Ψ^+). In other words, they describe classes of different diagnostic systems by varying only one parameter. A competence theory about diagnosis needs a more detailed framework to make more distinctions among different systems. Therefore, we have extended their framework to six parameters. As a result, more aspects of diagnostic systems can be used for classifying, choosing, reasoning about diagnostic systems. The extensions also give the possibility to describe more systems (see section 5). The next section describes our extended framework.

3 EXTENDED SPECTRUM OF LOGICAL DEFINITIONS

Section 3.1 informally describes the notion of diagnosis. This notion is captured in a formal description for diagnostic systems in section 3.2. The parameters of our definition are described in detail in section 3.3. The last section 3.4 explains some of the choices we made in our definition.

3.1 THE NOTION OF DIAGNOSIS

A diagnostic system computes the solutions for a diagnostic problem by using a behaviour description and the observed behaviour of the system under diagnosis. A diagnostic problem arises if there is a discrepancy between how a system (e.g. an artifact) behaves and how the system should behave, in other words the expected behaviour does not correspond with reality. The diagnostic task is to find out the cause of this discrepancy. The algorithm of the diagnostic system determines the possible causes of this discrepancy by using a behaviour model and the observed behaviour.

The general characteristics of a diagnostic system are the following. The computed solutions of a diagnostic problem represent an explanation for the observed behaviour. The diagnostic system exploits context information for solving the diagnostic problem. An explanation distinguishes *two types of observations*: it *covers* some observations, and it does *not contradict* other observations. The explanation is restricted to a *vocabulary* of special candidates that could be causes of a behaviour discrepancy (e.g. components) . Usually we are not interested in all possible explanations, but only in the most *reasonable* explanations for the current diagnostic problem. We also want to *represent* an explanation as a solution that the user can interpret. For example in medical domains, users are usually interested in the disease. They are not interested in all the current states of the parts of the patient's body.

Each of these aspects (written in italics) influences the particular notion of diagnosis that is realised in a given system. In the next section we capture these general characteristics of a diagnostic system in a formal definition.

3.2 THE DEFINITION OF DIAGNOSIS

In this section we give our definition of a diagnostic system. This definition contains a number of parameters. Different notions of diagnosis can be described by choosing values for these parameters. A diagnostic system is any system that satisfies the following definition for the input and output of the system.

Input: The behaviour model (BM) is a logical theory. As often in the literature, we assume that this is a Horn Clause theory. Furthermore, we extend this with preference conditions for clauses. A preference condition expresses when using the clause is preferred in respect to other clauses ([vanHarmelen & tenTeije, 1994]). The preferred explanations are the explanations where the preferences play the largest guiding role. There are two types of observed values. First, the observations that have to be explained (OBS), and second the contextual information (CXT) that helps to find the solution for the explainable observations (e.g. the preferences). CXT does not need to be explained by the diagnostic system.

Output: When given as input the behaviour model BM, a context CXT, and a set of observations OBS, a diagnostic system computes a set of solutions Sol such that:

$$
\begin{array}{lll}
(1) & \langle \Psi^+, \Psi^- \rangle = \underline{\Psi\text{-}mapping}(OBS) & and \\
(2) & Es = \{E| \quad BM \cup E \cup CXT \vdash_{\underline{cov}} \Psi^+ & and \\
(3) & \qquad BM \cup E \cup CXT \nvdash_{\underline{cov}} \neg\Psi^+ & and \\
(4) & \qquad BM \cup E \cup CXT \nvdash_{\underline{con}} \neg \Psi^- & and \\
(5) & \qquad E \subseteq \underline{Abducibles}\} & and \\
(6) & Selected(\underline{Selection\text{-}criterion}, Es, E') & and \\
(7) & \underline{Solution\text{-}form}(E', Sol) &
\end{array}
$$

Each of the underlined terms is one of the parameters in our definition. Varying one or more parameters means describing a different diagnostic system. The $\underline{\Psi\text{-}mapping}$ (rule 1) determines which observations must be explained (Ψ^+) and which observations need only not be contradicted (Ψ^-). E is an explanation for the observed behaviour by covering some observations (rule 2,3: $\underline{\vdash_{cov}}$), and not contradicting other observations (rule 4: $\underline{\nvdash_{con}}$). We write $\vdash_{cov}$ and $\nvdash_{con}$ as different symbols to emphasise that one is not necessarily the negation of the other, and that neither is necessarily the same as $\vdash$, but there is a relation between these explanation relations (see section 3.3.1). E is expressed in a particular vocabulary (rule 5: $\underline{Abducibles}$). We are interested in the most reasonable explanations, determined by $\underline{Selection\text{-}criterion}$ (rule 6). $\underline{Solution\text{-}form}$ (rule 7) produces Sol, the final output of the system and determines the representation of this solution.

Summarising, the input arguments are BM, CXT and OBS, the output argument are possible values of Sol and the parameters that describe the diagnostic system are the Ψ-*mapping*, the *Abducibles*, the cover relation $\vdash_{cov}$, the not-contradicting relation $\nvdash_{con}$,the *Selection-criterion* and the $Solution\text{-}form$. In the next section we discuss each parameter that is used in this definition.

3.3 THE PARAMETERS

In this section the parameters of the definition are described in more detail. First we describe the parameters which express "what" an explanation is. These are the parameters Ψ-*mapping*, *Abducibles*, *covering relation*, and *not-contradicting relation*. We continue with the selection parameter. This parameter determines what a "selected" explanation is. The last parameter description is the $Solution\text{-}form$ parameter. This parameter describes the "form" of a solution of the diagnostic problem. For every parameter we give the meaning, the form (e.g. relation, function) and some examples.

3.3.1 Parameters For An Explanation

• Ψ-Mapping

Meaning: The parameter Ψ-*mapping* maps the observed observations (OBS) onto two sets of observables: Ψ^- and Ψ^+. The explanation must cover the elements of Ψ^+. In other words, the strong explanation relation $\vdash_{cov}$ must hold for the elements in Ψ^+. The explanation must not contradict the elements in Ψ^-, thus the weak explanation relation $\nvdash_{con}$ must hold for these elements.

Form: The Ψ-*mapping* parameter is a function. The function has one argument OBS, the values of the observed observables. The output of the function is a tuple $\langle \Psi^+, \Psi^- \rangle$, whereby the Ψ^+ and Ψ^- are sets of observables. Notice (because Ψ-*mapping* is a function) we demand that $OBS = OBS' \rightarrow \langle \Psi^+, \Psi^- \rangle = \langle \Psi'^+, \Psi'^- \rangle$, but we do not demand the other way around: $OBS \neq OBS' \rightarrow \langle \Psi^+, \Psi^- \rangle \neq \langle \Psi'^+, \Psi'^- \rangle$. In other words: different OBS sets may lead to the same $\langle \Psi^+, \Psi^- \rangle$ division. We only require that all observations play a role in determining ψ^+ and ψ^- sets. This can be expressed by $OBS \subseteq (\psi^+ \cup \psi^-)$.

Examples:
An example of the knowledge that can be used for the construction of Ψ^+ and Ψ^- is the completeness of our knowledge about observable values. We can distinguish three kinds of observable values. (1) Our knowledge about the value is *complete* (e.g. $x = 5$). Such an observable belongs to Ψ^+; (2) our knowledge about the value is *incomplete* (e.g. $x \neq 5$): such an observable belongs to Ψ^-; (3) we have *no knowledge at all* about the value: such an observable belongs to neither Ψ^+ nor Ψ^-. The motivation for this definition is that we want to strongly explain the parameters that are completely known, and weakly explain (i.e. not contradict) the parameters that are partially known.

The Ψ-*mapping* can also be used to capture the differences between abduction based and consistency based diagnosis, as mentioned in [Console & Torasso, 1991]. system such as [Console & Torasso, 1990] the choice of Ψ-*mapping* is: $\Psi^+ = OBS$ and $\Psi^- = \emptyset$. In consistency based diagnostis such as [deKleer & Williams, 1987] the choice of Ψ-*mapping* is $\Psi^- = OBS$ and $\Psi^+ = \emptyset$.

Another example of a Ψ-*mapping* function is using knowledge about the observations in terms of normal or abnormal behaviour. In [Console & Torasso, 1991] it is motivated that it is reasonable to take Ψ^+ as the abnormal observations and Ψ^- as the normal observations. They also give the example of using the completeness of the behaviour model for the ψ-*mapping*. Completeness of the behaviour model for an observable

means that all the possible causes of the observable are represented in the behaviour model. If the observation has a complete model then it belongs to ψ^+ otherwise it belongs to ψ^-. Notice that Ψ-mapping instances often need extra knowledge for constructing the Ψ sets. In the examples above, we need extra knowledge about the observables, namely the kind of behaviour (normal or fault) and the completeness of its behaviour model.

• Abducibles

Meaning: The parameter *Abducibles* introduces the vocabulary of the explanation.

Form: The *Abducibles* is a subset of literals from the language of the behaviour model BM.

Examples: In abductive based systems the abducibles are often defined as the set of the initial nodes of a causal network [Console & Torasso, 1990]. An example in the consistency based stream is the *Abducibles* as a set of mode assignments [Dressler & Struss, 1992], where the *Abducibles* are all the fault and correct modes of each component of the system.

• Covering Relation $\vdash_{cov}$ And Not-Contradicting Relation $\nvdash_{con}$

Meaning: The two parameters *covering* relation and *not-contradicting* relation express the possible explanation-relations. These parameters are the basic notion of what counts as an explanation for the observed behaviour. The *covering* relation is the strong explaining relation ($\vdash_{cov}$), and the *not-contradicting* relation ($\nvdash_{con}$) expresses the weak explaining relation. The strong relation states the values in Ψ^+ *are required* as observations, and the weak relation states that the values in Ψ^- are *not ruled out* as observations.

Sometimes, diagnostic systems use only one of these relations. In such a case, one of the Ψ sets is defined as empty. For example in the pure consistency based systems only the not-contradict relation is used because $\Psi^+ = \emptyset$.

In the diagnosis literature these relation parameters are seldom subject of discussion. The first-order logic entailment is always taken for both the covering relation ($\vdash_{cov} \equiv \vdash$) and for the not-contradicting relation ($\nvdash_{con} \equiv \nvdash$). We think that these relations are interesting for classifying, comparing, choosing and reasoning about diagnostic systems and therefore they are appropriate as parameters for the spectrum. This is why we have introduced symbols distinct from $\vdash$. Similarly, we have introduced different symbols for $\vdash_{cov}$ and $\nvdash_{con}$ to emphasise that one need not necessarily be the negation of the other, but these explanation relations are related to each other. If both relations are defined then we require that any strongly explained observation should also be weakly explained.

Form: They are deduction relations. The relations have to be defined with respect to the fix form of the behaviour model that we have chosen: a Horn Clause theory extended with preference conditions on clauses.

Examples: As mentioned earlier, in current diagnostic systems these parameters are usually $\vdash$ and $\nvdash$. Alternative which can be useful are approximation of $\vdash$ or $\nvdash$. For example an incomplete but sound approximation of $\vdash$ or an complete but unsound approximation of $\nvdash$. Example of work in this area is [Cadoli & Schaerf, 1991]. Another example in this direction [Subramanian, 1987]. She constructs a calculus of weak and strong irrelevance. These could be examples of approximations of $\nvdash$. An example closer to the current diagnostic systems is a diagnostic system that computes cover-relations (causal-link chains) to a maximum depth.

By varying these relations we do not mean another implementation of the relation. We assume that the defined relation is implemented soundly and completely.

3.3.2 Parameter For A Selected Explanation

• Selected(Selection-criterion, Es, E)

Meaning: The selection parameter takes care of the choice of the most reasonable explanations among a set of explanations. Given the current situation CXT and observations OBS we are only interested in the most reasonable explanation, determined by *Selection-criterion*.

Form: The predicate *Selected(Selection-criterion, Es, E)* is true *iff* E is a selected explanation among the set of all explanations Es, using the predicate *Selection-criterion* as criterion. A *Selection-criterion* S is defined as a predicate $s(Es, E)$, which means E is selected among the elements of Es. Thus the following holds:

$$Selected(Selection\text{-}criterion, Es, E) \equiv Selection\text{-}criterion(Es, E)$$

More than one selected explanation of a set of explanations is allowed, in other words, given a value for Es the predicate *Selection-criterion* can hold for several E's. The *Selection-criterion* often uses information such as the behaviour model, the used explaining relations and context.

A *Selection-criterion* can be composed of several criteria. The reason for using *Selection-criterion* is because of the compositional character of the selection criterion. The *Selection-criterion* can be composed from constructors such as **comp** (composition), **and**,

or, **not** and **orelse** (if possible then x else y). Some examples of these constructor definitions are:

- composition of selection criteria:

$$Selected((C_1 \textbf{ comp } C_2), Es, E) \leftrightarrow \\ Es_{C_1} = \{E' | C_1(Es, E')\} \ \wedge \ C_2(Es_{C_1}, E)$$

- conjunction of selection criteria:

$$Selected((C_1 \textbf{ and } C_2), Es, E) \leftrightarrow \\ C_1(Es, E) \ \wedge \ C_2(Es, E)$$

- orelse composition of selection criteria

$$Selected((C_1 \textbf{ orelse } C_2), Es, E) \leftrightarrow \\ C_1(Es, E) \ \vee \ (\neg \exists E' | C_1(Es, E') \ \wedge \ C_2(Es, E))$$

Examples There are many examples of selection criteria in the diagnosis literature:

- Subset minimal with or without respect to special predicates (e.g. incompleteness assumptions in CHECK [Console & Torasso, 1990])
- The explanation must imply a set of diagnostic labels [Console & Torasso, 1990]
- The explanation must only contain the relevant causes [Peng & Reggia, 1986]
- The explanation must contain as few as possible fault mode assignments [deKleer & Williams, 1987].

An example of a composed *selection-criterion* is the selected explanation in CHECK. CHECK only selects explanations which imply some diagnostic labels. Furthermore, CHECK prefers stronly confirmed explanations over weakly confirmed explanations which are again preferred over subset minimal explanations:

E-label **and** *(strongly-confirmed* **orelse** *weakly-confirmed* **orelse** *subset-minimal)*

3.3.3 Parameter For The Solution

• Solution-form(E, Sol)

Meaning: $Solution\text{-}form$ represents the explanation of the observed behaviour in such a way that it is appropriate for the user of the diagnostic system.

Form: $Solution\text{-}form$ is a relation, so one explanation E may relate with more then one Sol. Notice that a diagnostic system delivers a set of set of Sol's: we can have multiple explanations E, and each explanation E may yield different solution forms Sol.

Examples: CHECK recognizes the need for diagnostic labels. A diagnostic label is an abbreviation of a set of states. This is for example useful in the medical domain: doctors use diseases as diagnosis instead of the set of states that explains the complaints. This is exactly what $Solution\text{-}form$ encodes; the explanation in appropriate form for the user. Another example is expressing the solution as the path that is used for explaining the observed behaviour. This example illustrates that sometimes one explanation can result in more solutions, because there may be several ways that could be used for finding the same explanation.

3.4 DISCUSSION

In this section we answer three obvious questions.

Why is the behaviour model not a parameter? The form of the behaviour model BM is fixed in our framework. It is a Horn Clause theory extended with preference conditions on clauses.

There are three reasons for not taking the behaviour model as a parameter of the definition. First if the structure of the behaviour model differs (e.g. default rules versus predicate calculus sentences) then it is very hard to compare these diagnostic systems. The second reason is that it is possible to express behaviour models of several diagnostic systems in the form of the proposed behaviour model. This means that we are not too restrictive, because we are able to describe several divergent diagnostic systems in the extended spectrum. The third reason is that we get a less complex framework by choosing the behaviour model as a fixed form. We prefer to test our ideas on a framework that is not too complex, and there is still the possibility of extending the framework with the behaviour model form as a parameter.

At this moment we are not able to express diagnostic systems, which use a behaviour model that cannot be translated to a Horn Clause theory.

Why do we not use a set of components as input argument? The consistency based community uses the set of components as input argument. We do not use this set explicitly, because we can encode this by the *Selection-criterion* and the *Abducibles* parameters as follows:

We define the parameter *Abducibles* as the set of fault modes and ok modes for every component. As a result, the explanation contains only mode assignments. We then demand in the *Selection-criterion* that for every component there must be exactly one mode assignment in the explanation. In this way we encode the usual use of the set of components, and there is no reason to demand the set of components as an input

argument.

Why is there not a difference between initial and additional observations? In our framework, we define the competence of diagnostic systems. The difference between initial and additional observations is of an algorithmic nature, whereas we focus on the declarative description of diagnostic systems. This excludes that there should be a difference in observation sequence.

4 EXAMPLES: DESCRIPTIONS OF DIAGNOSTIC SYSTEMS

We have described several diagnostic systems in the proposed extended spectrum. They are: a pure abductive system ([Console & Torasso, 1990]), an abductive system with preferences ([Eshelman, 1988]), an abductive system with abstractions ([Console & Dupré, 1992]), a set-covering system ([Peng & Reggia, 1986]), a pure consistency based system ([deKleer & Williams, 1987]), a consistency based system with abstractions ([Mozetic, 1991]), a consistency based system with fault modes ([Struss, 1990]), and a consistency based system with preferences ([Dressler & Struss, 1992]). Many of these systems cannot be described in the original spectrum of Console and Torasso.

This section contains two examples of diagnostic systems descriptions. We choose the abductive hierarchical system of [Console & Dupré, 1992], because of the specific behaviour model that this system uses. This behaviour model is translated into our fixed form. They also use a complex preferred explanation, which is expressible in our *Selection-criterion*. Secondly we choose the system from [Struss, 1990], because it is a consistency based system in contrast to the above system, which belongs to the abductive stream.

The two following sections are structured as follows: first we give a short description of the basic ideas of the diagnostic system, then we describe the system in our spectrum. This contains the translation of the behaviour model of the diagnostic system to our representation of the behaviour model, and the instantiations of every spectrum parameter.

4.1 HIERARCHICAL ABDUCTIVE DIAGNOSIS

4.1.1 Introduction

The system from [Console & Dupré, 1992] uses two kinds of theories: explanation theories and abstraction theories. The explanation theories contain causal links, and the abstraction theories contain abstraction relations between symbols that occur in the explanation theories. In this system, the negative observations have to be consistent with both types of theories. The positive observations have to be covered in these theories, but the idea is that the observations have to be covered with causes at the right level of abstraction. The level of detail of the observations guides the level of detail of the explanation. The basic idea of this system is that low level assumptions are only allowed if there are specific observations related to them or if high-level explanations cannot be found. Another demand in this system is to express the explanation as "simple" as possible, i.e. to avoid redundant high-level assumptions.

4.1.2 Translation Of The Behaviour Model

As explained above, the behaviour model of this system consists of a number of explanation theories and abstraction theories. These are Horn Clause theories, so they already have the right form for our BM. This BM will be extended with preference conditions, because the system prefers the use of explanation clauses over abstraction clauses. We can express this by giving false preference conditions to the abstraction clauses and true preference conditions to the explanation clauses.

4.1.3 Parameters

Ψ-mapping: The division of observations is based on positive or negative observation of the observables. The Ψ^+ contains the positively observed observations and the Ψ^- contains the negatively observed observations. The system ignores the observables, which are neither positively nor negatively observed.

Abducibles: The explanation vocabulary consists of the initial nodes of the explanation theories. In our terms, these are the initial nodes of the preferred clauses in BM .

Cover relation: As usual, $\vdash$ of first order logic.

Not-contradict relation: As usual, $\nvdash$ of first order logic.

Selection-criterion: In this system the selected explanations are the explanations that use preferred clauses instead of unpreferred clauses, if possible. In other words the system uses only abstraction relations if there is no other way to solve the diagnostic problem. A clause is preferred when its preference condition is true. All explainable relations have true preference conditions, so they are preferred in an explanation. The formal definition of preferences is described in [vanHarmelen & tenTeije, 1994].

Solution-form(E, Sol): This system represents E as simple as possible, which means roughly that explana-

tion with fewer causes are better. This is described formally as follows, $Solution\text{-}form(E, Sol)$ *iff*:

$$\begin{array}{ll} explanation(Sol) & \wedge \\ Sol \cup BM \cup CXT \vdash E & \wedge \\ Selected\text{-}explanation(E) & \wedge \\ \neg\exists Sol' : \quad Sol' \subset Sol \ \wedge \ explanation(Sol') \wedge & \\ \qquad Sol' \cup BM \cup CXT \vdash E & \end{array}$$

whereby an $explanation(x)$ is true *iff* for x both explaining relations hold, and x is a subset of $Abducibles$. Notice that E is a selected explanation, while Sol is an explanation that is not necessarily selected.

4.2 CONSISTENCY BASED DIAGNOSTIC SYSTEM WITH FAULT MODES

4.2.1 Introduction

The system from [Struss, 1990] is a consistency based diagnostic system like GDE [deKleer & Williams, 1987]. This kind of diagnostic systems uses a structure description and behaviour models of the components of the system for explaining the unexpected behaviour. This is done by an explanation, which represents which components work correctly and which are not correct. Here we describe a diagnostic system that also uses fault modes for explaining the unexpected behaviour. In such systems the explanation represents not only which components are correct and incorrect, but the incorrect components are also represented by their fault modes. The explanation is a set with an ok -mode or a specific fault mode for every component in the system. Because the system does consistency based diagnosis, this set must be consistent with the fault models, behaviour component models, structure model and the context. The selected explanations are the explanations with the fewest fault modes. The user of the system gets only the fault modes as solution for the diagnostic problem.

4.2.2 Translation Of The Behaviour Model

The fault models and correct behaviour models of the components and the structure model of the system together constitute the BM.

4.2.3 Parameters

Ψ-mapping: In consistency based systems are all the observations are only weakly explained. So: $\Psi^- = OBS$ and $\Psi^+ = \emptyset$.

Abducibles: An explanation is a set of assignments (fault modes or correct mode) of the components of the system. The *Abducible* set is

$$\{ok(c), fm_i(c), \neg ok(c), \neg fm_i(c) | c \in COMP\}$$

.

Explain relation: The set Ψ^+ is by definition empty, so the cover relation is not relevant here.

Not-contradict relation: As usual, $\nvdash$ of first order logic.

Selection-criterion: This system selects those explanations with a maximal set of ok components and every component has a mode assignment. We can describe this as $Selected(max\text{-}ok$ **and** $all\text{-}comps, Es, E)$, where:

$max\text{-}ok(Es, E)$ *iff*

$$\begin{array}{ll} E \in Es \wedge & \\ \neg\exists E' : \quad E' \in Es \wedge & \\ \qquad \|\{c | c \in COMP \wedge ok(c) \in E'\}\| > & \\ \qquad \|\{c | c \in COMP \wedge ok(c) \in E\}\| & \end{array}$$

$all\text{-}comps(Es, E)$ *iff*

$$\begin{array}{l} E \in Es \wedge \\ \forall c \in COMP : \exists! \alpha \in \{ok, fm_i\} : \alpha(c) \in E \end{array}$$

Solution-form(Sol, E): The solution contains only the fault modes of the explanation. Thus: $Sol = \{fm_i(c) | fm_i(c) \in E\}$

5 ADVANTAGES OVER THE ORIGINAL SPECTRUM

5.1 THE ORIGINAL SPECTRUM

The spectrum of Console and Torasso is a framework for the formal description of several diagnostic systems. Abduction based, consistency based and association rule based systems can be easily formulated in their framework.

The input of systems in their spectrum is the system description (SD), context CXT and the observations OBS. The system description consists of the behaviour model BM (a Horn Clause theory) and a set of components $COMP$. In this diagnostic framework an explanation E is defined as a set of mode assignments for each component, where a particular set of observables has to be covered by the explanation, and another set of observables has to be consistent with the explanation. This second set is computed in a fixed way on the basis of the first set. The abducibles that can be used in the explanation are the behavioral modes of the components. The set of abducible predicates is derived from BM in a fixed way. Symbols are abducible *iff* they do not appear in the head of any clause in BM. As a result, BM determines for every predicate symbol whether it is an abductive or a

non-abductive predicate symbol. The selection criterion is also fixed: a selected explanation is a minimal set of faulty components. The single parameter of this framework is the set of observables that must be covered (Ψ^+). By varying this parameter, the spectrum ranges from $\Psi^+ = OBS$ (abductive) to $\Psi^+ = empty$ (consistency-based).

Given as input the system description $\langle BM, COMP\rangle$, a context CXT, and a set of observations OBS, a diagnostic system in the framework of Console and Torasso computes a set of explanations E such that:

$$
\begin{array}{lll}
(1) & \Psi^+ \subseteq OBS & \\
(2) & \Psi^- = & \{\neg f(x) | f(y) \in OBS, \\
 & & \text{for each } f, \text{for each admissible } x \neq y\} \\
(3) & Es = & \{E| \quad BM \cup E \cup CXT \vdash \Psi^+ \ \wedge \\
(4) & & \quad BM \cup E \cup CXT \nvdash \neg\Psi^+ \wedge \\
(5) & & \quad BM \cup E \cup CXT \nvdash \neg\Psi^- \wedge \\
(6) & & \quad \forall c \in COMP: \ \exists! \alpha : \alpha(c) \in E \\
 & & \quad \wedge \ \alpha \in abducibles(BM)\} \\
(7) & E' \in Es & \\
(8) & \neg\exists E'' \in Es : faulty(E'') \subset faulty(E') &
\end{array}
$$

where

$$faulty(E^*) = E^* \setminus \{ok(c)| \ c \in COMP\}$$

5.2 ADVANTAGES OF OUR EXTENSIONS

The original framework can be described in our extended framework by choosing appropriate values for the six parameters. This shows that the extended framework is at least as general as the original one.

The original spectrum describes a diagnostic system by varying a single parameter. However, our extended spectrum describes a diagnostic system by six parameters. In more detail we can say that our Ψ-mapping function is more flexible. We can express the original subdivision of Ψ^- and Ψ^+ in our Ψ-mapping parameter. In the original spectrum the Ψ^- is fixed. Our more flexible Ψ-*mapping* is for example useful if we want to cover the abnormal observations and not contradict the normal observations. This example of the Ψ-*mapping* is not possible in the original spectrum, but is used in actual diagnostic systems.

The E is a subset of the abducibles in our extended spectrum. In the original spectrum this is restricted, namely E has to be a mode assignment for every component in the set of components. There are diagnostic systems which do not demand that every component needs a mode assignment in the explanation. This means that the original spectrum is too strict in this aspect.

In our spectrum the abducibles are also variable. In the original spectrum the abducibles are not variable, but derived from BM. The hierarchical system from [Console & Dupré, 1992] can not be described in the original system because the abducible set would be derived incorrectly.

In the extended spectrum we try to represent the notion of covering ($\vdash_{cov}$) and not-contradicting ($\nvdash_{con}$) the observations instead of the specific relations that are often used in the diagnostic systems, namely $\vdash$ and $\nvdash$. The original spectrum uses only these standard relations.

The extensions of the spectrum have two effects for the systems that can be expressed in the spectrum. First our extended spectrum delivers **more distinctions** among system descriptions. There are more aspects of diagnostic systems that we can use for choosing, comparing and reasoning about them. For example, our extended spectrum enables choosing of a diagnostic system based on the minimality criterion, or the form of the explanation, but in the original spectrum this is impossible. A second effect is that **more systems** fit in the extended spectrum. Examples of diagnostic systems, which are expressible in the extended spectrum, but not in the original spectrum are systems that use another minimality criterion than subset-minimal in fault components, and systems that use another Ψ division.

6 RELATED WORK

This section is a short description of other papers that proposed a diagnostic theory, but which we did not take as starting point of our framework as motivated below.

[Benjamins, 1993]: This modeling framework is based on the notions of problem solving methods and tasks, where problem solving methods are viewed as ways to decompose tasks into subtasks. The framework includes the notion of method suitability in order to specify the conditions under which methods can be applied to tasks. Benjamins only partially represent the competence of a diagnostic system. The full competence of a system can only be derived in retrospect from the composition of the system from subcomponents. We want to abstract from the composition of the system, and only reason about the competence of the system as a whole, but in Benjamins' system this composition is required for the derivation of the competence. A second difference with our definitions is that Benjamins' framework does not only capture the input/output relation (competence) but also the algorithm of the system. Finally, the work of Benjamins' has until now not been sufficently formalised to use it as the basis for automated reasoning, which is one of our goals.

[Böttcher *et al.*, 1991]: GenDE is a framework for algorithms in consistency based diagnosis. In this framework a diagnostic system is composed of several modules. The basic modules are: the predictor, the candidate proposer, and the strategist. Each module can be composed of other submodules. The three basic components describe the algorithmic aspects of consistency based system nicely. Examples as GDE, GDE with hierarchical knowledge and GDE with fault modes fit in the framework. As we mentioned earlier we are interested in an explicit representation of the input/output relation. In the GenDE framework this relation is hidden in several modules. GenDE is not the most appropriate framework for our goal because we are not focussing on algorithms and the input/output relation is not separately described in the framework. Another disavantage for our goal is the nature of the systems that can be described in GenDE. GenDE is more aimed at consistency based systems.

[Leitch *et al.*, 1993]: The framework of Leitch is a mapping from problem space to solution space for model based diagnostic systems. This diagnostic theory captures the criteria that determine which specification of a diagnostic system is appropriate in a particular case. The mapping consists of three steps. The first step is to determine the possible strategies based on problem requirements for task, faults and models. The second step is the choice of the prediction method. This choice is based on the possible strategies, the fault requirements and model requirements. The third step is the selection of the candidate generator. This selection is based on the fault requirements. Again, in this framework there is no description of the input/output relation. Leitch's work corresponds more with the work of GenDE, because it distinghuishes the same three basic modules, namely the strategist, the predictor, and the generator. GenDE emphasizes the specification of these modules and Leitch's emphasizes the criteria that determine the appropriate modules. Notice that [Benjamins, 1993] also uses several criteria for determining the methods suitability. The goal of Leitch's theory is to specify model based diagnostic systems by the three modules, in contrast with the theory of Console and Torasso, that specifies the diagnostic system by its input/output relation. In Leitch's theory the input/output realation is implicite in the three modules. Furthermore, Leitch's theory does not include pure abduction. The theory is intended for consistency based systems. The theory does distinguish however the consistency based systems with and without faultmodes.

7 SUMMARY

Our motivation for developing a framework for diagnostic reasoning is to create a basis for comparing, choosing, and reasoning about diagnostic systems. The proposed framework captures many different notions of diagnosis. Using six parameters the framework describes a diagnostic system in the range of pure abductive systems to pure consistency based systems. The concepts that are described in the framework are: (1) an explanation of the observed behaviour by four parameters: Ψ*-mapping, Abducibles, covering relation, not-contradict relation*; (2) the selection of an explanation by the *Selection-criterion* parameter; and (3) the final form of the solution to the diagnostic problem by the *Solution-form* parameter. An explanation is based on two kinds of explanation relations, a strong *covering relation* that must hold for a set observables Ψ^+ and a weaker *not contradicting relation* that must hold for another set of observables Ψ^-. The construction of these two sets of observables is based on the observed behaviour and context of the diagnostic problem. This construction (represented by the parameter Ψ*-mapping*) is one of the characteristics of a diagnostic system. Another characteristic is the vocabulary (*Abducibles*) that is used for the explanation. A selected explanation is selected among all the possible explanations using the specified *Selection-criterion*. The final solution of the diagnostic system is expressed in appropriate terms for the user. This charateristic is expressed in the parameter *Solution-form*. The framework is based on the spectrum of Console and Torasso. The extended spectrum differs from the original one at two points: the detail of description and the number of diagnostic systems that can be described. The result of the extensions is that we are able to express **more** systems and to express a diagnostic system in **more detail**.

Acknowledgement

We are grateful to Gertjan van Heijst and Manfred Aben for their comments on an earlier version of this paper. Our work has benefitted from discussions with Guus Schreiber and Bob Wielinga.

References

[Benjamins, 1993] V. R. Benjamins. *Problem Solving Methods for Diagnosis*. PhD thesis, University of Amsterdam, Amsterdam, The Netherlands, June 1993.

[Böttcher *et al.*, 1991] C. Böttcher, O. Dressler, H. Freitag, and M.Montagand P. Struss. Architectural design and specification of gende. Technical report, December 1991. Project Behaviour, memo 12-91.

[Cadoli & Schaerf, 1991] M. Cadoli and M. Schaerf. Approximate entailment. In *Proceedings of 2nd Italian*

*AI Conference AI*IA'91*, pages 68–77. Springer Verlag, Lecture Notes in AI, No. 549, 1991.

[Console & Dupré, 1992] L. Console and D. Theseider Dupré. Choices in abductive reasoning with abstraction axioms. In G. Lakemeijer, editor, *Proceedings of the workshop on foundations of knowledge representation*, Vienna, August 1992. ECCAI.

[Console & Torasso, 1990] L. Console and P. Torasso. Hypothetical reasoning in causal models. *Int. J. of Intelligent Systems*, 5(1):83–124, 1990.

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

[deKleer & Williams, 1987] J.H. de Kleer and B.C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32:97–130, 1987.

[Dressler & Struss, 1992] O. Dressler and P. Struss. Back to defaults: characterising and computing diagnoses as coherent assumption sets. In *Proceedings of ECAI'92*, pages 719–723, Vienna, Austria, August 1992.

[Eshelman, 1988] L. Eshelman. MOLE: A knowledge-acquisition tool for cover-and-differentiate systems. In S. Marcus, editor, *Automating Knowledge Acquisition for Expert Systems*, pages 37–80. Kluwer, Boston, 1988.

[Leitch *et al.*, 1993] R.R. Leitch, M.J. Chantler, Q. Shen, and G.M. Coghill. A preliminary specification methodology for model based diagnosis. In *Fourth International Workshop on Principles of Diagnosis*, Aberystwyth in Wales, September 1993.

[Mozetic, 1991] Igor Mozetic. Hierarchical model-based diagnosis. *Int. J. of Man-Machine Studies*, 35(3):329–362, 1991.

[Peng & Reggia, 1986] Y. Peng and J. A. Reggia. Plausibility of diagnostic hypotheses: The nature of simplicity. In *Proc. AAAI-86*, pages 140–145, Philadelphia, PA, August 1986. AAAI.

[Poole, 1989] D. Poole. Normality and faults in logic-based diagnosis. In *Proc 11th. IJCAI*, pages 1304–1310, Detroit, 1989.

[Reiter, 1987] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–96, 1987.

[Struss, 1990] P. Struss. Problems of interval-based qualitative reasoning. In D. S. Weld and J. H. de Kleer, editors, *Readings in Qualitative Reasoning about Physical Systems*. Morgan Kaufmann, San Mateo, California, 1990.

[Subramanian, 1987] D. Subramanian. The relevance of irrelevance. In *Proceedings of IJCAI'87*, page some pages, Milan, August 1987.

[vanHarmelen & tenTeije, 1994] F. van Harmelen and A. ten Teije. Using domain knowledge to select solutions in abductive diagnosis. In *Proceedings of ECAI'94*, pages 652–656, Amsterdam, August 1994.