A Formal Framework for Representing Diagnosis Strategies in Model—Based Diagnosis Systems

Wolfgang Nejdl, Peter Prohlich and Michael Schroeder
University of Hannover, Lange Laube 3, 30177 Hannover, Germany
e-mail: nejcll@kbs.uni-hannover.de

Abstract

Recent work has pointed out that diagnosis strategies are a necessary tool for the diagnosis of complex systems. Nevertheless, though current diagnosis systems are able to use explicit system models, their representation of diagnosis strategies is only implicit. In this paper we introduce a formal meta language to express strategic knowledge in an explicit way, This language is sufficient to formalize all strategies introduced in previous work, and extends previous diagnosis strategies by the integration of empirical knowledge and by explicit statements about dependencies between actions. We provide a declarative semantics for this language and an architecture for implementation.

1 Introduction

In order to handle the complexity of large-scale diagnosis we have to use more than one system model and see diagnosis as a dynamic process controlled by diagnosis assumptions made explicit as working hypotheses (as formalized first by Struss in [8]). In the same spirit Boettcher and Dressier ([I], [2]) developed a catalogue of diagnosis strategies and provided an intuitive semantics and an ATMS-based implementation for these strategies. The disadvantage of their approach is that they use a static set of strategies which is coded into the diagnosis algorithm. Missing a declarative semantics for these diagnosis strategies independent of a particular implementation makes the definition of new or application-specific strategies more difficult than it should be.

In this paper we extend this approach by introducing a formal meta-language for the definition of diagnosis strategies. This language makes strategies explicit and allows to define strategies specific to an application similar to defining system models. So our framework extends model-based diagnosis in the sense that not only the behavior of the system but also the strategic knowledge about the system model is represented explicitly.

2 Working Hypotheses

We consider a system described by a set of formulas *SD* in a language *C*. An observation *OBS* of the system *SD* is a finite set of formulas in $\mathcal{L}$. $\mathcal{L}$ is a first order language with equality. For simplicity we postulate that $\mathcal{L}$ contains no function symbols with variable interpretation. Functions with standard interpretation like mathematical operators etc. are allowed. By $ATOMS$ we denote the set of atoms of $\mathcal{L}$. Our theory does not depend on a particular diagnosis definition. We encapsulate the underlying diagnosis concept by a function *diag*, which maps a theory T to a set of diagnoses $\mathcal{D}$:

$$diag(T) := \begin{cases} \emptyset, & \text{if } T \text{ contradictory} \\ \text{a set of diagnoses} \\ \mathcal{D} = \{D_1, \ldots D_n\}, & \text{otherwise} \end{cases}$$

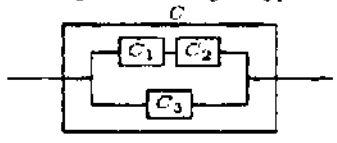
This general definition allows for a wide range of diagnosis concepts like minimal diagnoses [7], most probable diagnoses [5], preferred diagnoses [4] and others.

Struss introduced the concept of working hypotheses into model–based diagnosis in order to make diagnostic assumptions explicit [8]. The diagnostic assumptions are necessary in the diagnostic process for evaluating hierarchies, using simplified behavioral models, focusing on a particular kind of diagnoses, etc.

Definition 2.1 *Working Hypothesis*
Let $WHYP \subseteq ATOMS(\mathcal{L})$ be a set of atoms. We call each ground atom $wh \in WHYP$ a Working Hypothesis.

We name the general set of working hypotheses $WHYP$, a subset WH and elements of theses sets wh. Working Hypotheses can be used to represent multiple models of the system within one system description as is shown by the following example.

Example 2.2 *Use of Working Hypotheses*



A component C consists of the subcomponents $C_1, \ldots C_3$. The working hypothesis $refine(C)$ can be used to switch between the abstract model of C and the detailed model of C in which its subcomponents $C_1, \ldots C_3$ are visible. In the system description SD for some device containing C the behavior of C is modeled depending on $refine(C)$:

$$\neg refine(C) \rightarrow \textit{Rules for the abstract model of } C$$
$$refine(C) \rightarrow \textit{Rules for the detailed model of } C$$

Now if we want to compute the diagnoses for the system based on the detailed model of C we add refine(C) to the system description, i.e. we compute diag(SD ∪ OBS ∪ {refine(C)}).

In general, to compute the diagnoses under a set of working hypotheses WH we add WH to the system description. Additionally we add $\{\neg wh | \ wh \in WHYP \backslash WH\}$, i.e. we make sure that WH is exactly the set of working hypotheses which are true in the system model.

Definition 2.3 *Diagnosis under a Set WH of Working Hypotheses*
Consider a system described by $\langle SD, COMPS, OBS\rangle$, where SD is the system description, COMPS is the set of the system components and OBS is a set of observations. Let WH be a set of Working Hypotheses. Let $\overline{WH} = WHYP \backslash WH$. Then the set of diagnoses under working hypotheses WH, called $diag_{WH}$ is defined as follows:

$$diag_{WH}(SD \cup OBS) := \{D \cup WH \mid D \in diag(\ SD \cup OBS \cup WH \cup \{\neg wh | \ \ wh \in \overline{WH}\})\}$$

We include the set WH itself in the diagnosis, so that it is obvious from the diagnoses to which system model they belong. So $diag_{WH}$ provides a valid diagnosis concept as $SD \cup OBS \cup D$ is consistent, where $D \in diag_{WH}(SD \cup OBS)$.

Working hypotheses are an important concept for making the current diagnosis assumptions explicit. But the selection of suitable working hypotheses for a given situation is implicit in current diagnosis systems. In the next section we introduce a language that makes the knowledge for selecting the right hypotheses explicit and thus provides a flexible and declarative way of specifying strategic knowledge for the diagnostic process.

3 A Formal Language for Strategies

3.1 Preliminary Considerations

Diagnosis strategies control the diagnostic process by specifying which diagnosis assumptions should be used in a given situation. The state of the diagnostic process manifests itself in the current set of possible diagnoses. Thus the specification of a diagnosis strategy consists of

- a property of the current set of diagnoses, characterizing a certain situation that can occur during the diagnostic process
- an assumption or action modeled by a working hypothesis that is suitable for handling that situation

Example 3.1 *Diagnosis Strategy*
Consider an abstract component C as described in example 2.2. By default, we only use the abstract model of this component for diagnosis, i.e. —refine(G) is used as working hypothesis. The detailed model is only used when C is identified as faulty. This can be captured by the following rule:

> *If an abstract component C occurs in all diagnoses, activate a more detailed model for C making its subcomponents visible to the diagnostic process.*

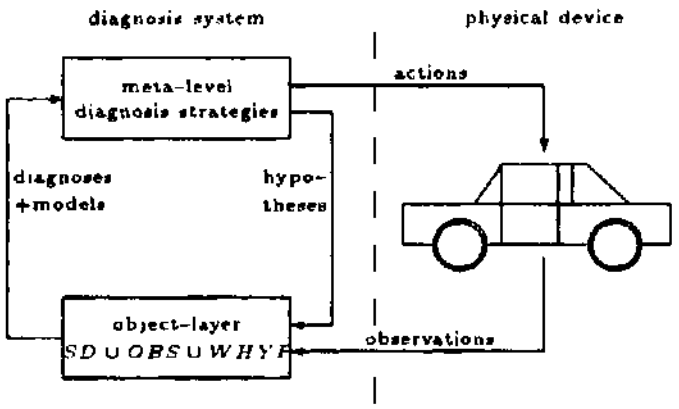


Figure 1: Diagnosis System with a Meta Level

In order to check if a given strategy should be applied we have to evaluate a condition on the current set of diagnoses. Such a condition cannot be modeled as part of the system description. So Boettcher and Dressier implement the check of these conditions as part of the diagnosis system.

However, the only flexible way of evaluating these conditions is by introducing a meta-level in the diagnosis system as shown in figure 1. So the strategies are defined on the meta level and they are evaluated using the knowledge obtained so far during the diagnostic process which is represented by the current possible diagnoses.

So *diagwfi(SD* U *OBS)* as defined in the previous section is the information on which the decisions on the meta-level are based. For some strategies it is not sufficient to have information about the faulty components only. For example, one necessary precondition for proposing a measurement point is that the value at that point is not known. To evaluate this condition we need to extend the diagnoses by the values predicted in the corresponding system models. So we postulate that *diagwH(SD UOBS)* contains all predicates needed by the preconditions of the diagnosis strategies.

3.2 The Meta Language

The language *Cstrat* for defining diagnosis strategies defines modal logic operators specifying properties of the current diagnoses as well as for proposing working hypotheses. Before we give a formal definition of the language we motivate the need for these modal operators informally.

Modal Operators for Characterizing the Current State of the Diagnostic Process:

As already stated the preconditions for the application of diagnosis strategies are statements about the current set of possible diagnoses. The atomic statements in these conditions are:

- a property *p(x)* is true under all possible diagnoses, or
- a property *p(x)* is true under at least one possible diagnosis.

These statements can be formalized using the usual S5 modal operators (□ for knowledge, ◊ for belief):

$\Box p$ p is true under all diagnoses of $SD \cup OBS \cup WH$.

$\Diamond p$ p is true under at least one diagnosis of $SD \cup OBS \cup WH$.

Modal Operators for Proposing Working Hypotheses

Strategy formulas specify which working hypotheses should be assumed in a given situation. This is achieved by the following (informally described) modal operators:

$\blacksquare\Box wh$ wh is a necessary working hypotheses in the current situation, i.e. the diagnostic process cannot be continued without assuming wh.

$\blacklozenge\Box wh$ wh is a possible (allowed) working hypotheses in the current situation.

Actions can also be expressed in this approach by using a restricted form of procedural attachment in the system description. For example, we have a working hypotheses mcasure(x) for proposing a measurement. The measurement itself is then implemented by a procedure get_value, which executes the measurement and remembers the value. The combination of the two concepts is modeled by a rule

$$\forall x : \forall v : measure(x) \rightarrow (\mathtt{get\_value}(x, v) \leftrightarrow val(x, v))$$

in the system description. Thus, from the logical viewpoint, the only effect of the predicate get^value is that it tests if v is the value of x. Our language allows to explicitly represent dependencies between actions, eg it is possible to express that action a is to be preferred over action b if both are possible (see section 5).

Syntax of Cstrat

Besides the modal operators Lstrat contains the usual logical connectives and quantifiers. In the following definitions we only consider V,—>,A since this is already a complete set.

Definition 3.2 *$\mathcal{L}_{Strat}$-Formula*

1. *Let L be a formula in the language $\mathcal{L}$. Then $\Box L$ and $\Diamond L$ are $\mathcal{L}_{Strat}$-Formulas.*
2. *Let S be a $\mathcal{L}_{Strat}$-Formula, then $\blacklozenge S$ and $\blacksquare S$ are $\mathcal{L}_{Strat}$-Formulas.*
3. *Let S_1, S_2 be $\mathcal{L}_{Strat}$-Formulas. Then also $\neg S_1$ and $S_1 \wedge S_2$ are $\mathcal{L}_{Strat}$-Formulas.*
4. *Let v_1, v_2 be variables. Then $v_1 = v_2$ is an $\mathcal{L}_{Strat}$-Formula*
5. *Let v be a variable, S an $\mathcal{L}_{Strat}$-Formula. Then $\forall v : S$ is an $\mathcal{L}_{Strat}$-Formula.*
6. *Nothing else is an $\mathcal{L}_{Strat}$-Formula.*

Note, that $\mathcal{L}_{Strat}$ has the same predicate symbols and constants as the system description language $\mathcal{L}$. The variables denote objects of $\mathcal{L}$. This will be ensured in the formal semantics presented in the next section. The following is an example for a strategy formula:

Example 3.3 Lstrat-Formula for Structural Refinement

For component C the strategy "Structural Refinement." can be expressed by the strategy formula

$$\Box ab(C) \rightarrow \blacksquare\Box refine(C).$$

By introducing a predicate refinablc(c) in the system description, which is true for all components that have subcomponents, we can generalize this rule:

$$\forall c : (\Box refinable(c) \wedge \Box ab(c) \rightarrow \blacksquare\Box refine(c))$$

More examples for the formalization of diagnosis strategies can be found in section 5. In the next section we define the declarative semantics for the strategy language.

4 Declarative Semantics for Strategies

We consider a diagnosis problem that is now described by

(SD, STRAT, COMPS, DBS)

where SD is the system description, ST RAT is a set of L$f rat-Formulas, COMPS is a set of components and OBS is a set of observations.

Diagnosis strategies are used to guide the diagnostic process. The semantics introduced in this section answers two questions:

- What are the possible sequences of diagnostic decisions implied by the given diagnosis strategies, i.e. which diagnostic process is consistent with respect to the given strategies STRAT (and with SD, CO MPS, OBS)?
- Which diagnoses are the result of such a consistent diagnostic process?

4.1 Characterization of the Diagnostic Process

The state of the diagnostic process can be characterized by a set of working hypotheses. Then, given SDUOBS the diagnoses can be inferred by applying diagwH-

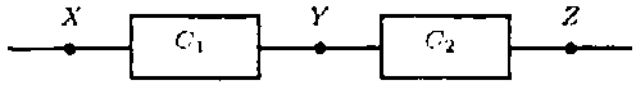
The diagnostic process as a whole can be characterized by specifying which states can be adopted or which transitions between states are considered. We will therefore define a state transition relation as follows:

Definition 4.1 State Transition Relation

A State Transition Relation is a binary relation 7v C 2 WHYP x yWHYP t e a relation among sets of working hypotheses.

In the next example we show how a diagnostic process can be encoded by a state transition relation:

Example 4.2 *Diagnostic Process*



Suppose we have two components C_1 and C_2, each having a structure as used in example 2.2. For this system we use strategies for structural refinement, and a measurement strategy that proposes Y as a measurement

point if its value t$ unknown. Having made some observation we start the diagnosis in the most abstract model without assuming any hypotheses. A look at the resulting diagnoses reveals that we do not know the value of Y and so we cannot say whether C\ or C2 [ls] faulty. So we measure the value of Y, i.e. we adopt the hypothesis measure(Y) and compute the diagnoses again. Now we identify C_2 as the faulty component. Therefore we look at a more detailed model of $_{C2}$ to find out which subcomponent of C_2 caused the error. So we additionally adopt the hypothesis refine{C>2)- After recomputing the diagnoses we know that C_{22} is the faulty component. This diagnostic process can be characterized by the following state transition relation R:

$$\mathcal{R} = \{(\emptyset, \{measure(Y)\}),\\ (\{measure(Y)\}, \{measure(Y), refine(C_2)\}),\\ (\{measure(Y), refine(C_2)\},\\ \{measure(Y), refine(C_2)\})\}$$

The first element of R denotes that in the state described by the empty set of working hypotheses we adopted the hypothesis measure(Y) which caused a measurement. After measuring we additionally considered the hypothesis refine(G2)- The termination of the diagnostic process in the state {measure(Y), refine{C2)} is modeled by the cycle in that state.

The course of the diagnostic process need not be linear. If we had also considered refining component C_1 instead of making the measurement, we would characterize the diagnostic process by

$$\mathcal{R}' = \mathcal{R} \cup \{(\emptyset, \{refine(C_1)\}),\\ (\{refine(C_1)\}, \{refine(C_1)\})\}$$

So, the state transition relation is just an encoding of the working hypotheses we used in each step of the diagnostic process and is influenced by the diagnoses found and the strategies given. In the declarative semantics we judge whether or not a diagnostic process (represented by a state transition relation) is correct wrt a diagnosis problem characterized by (SD, ST RAT, C O M PS, OBS).

The procedural semantics which is only briefly discussed in this paper computes correct diagnostic processes based on this semantics. The user of the diagnosis system only has to provide the system description and the strategies.

In order to check if the decisions made during the diagnostic process are consistent with the given diagnosis problem, we define how strategies can be interpreted as statements about the diagnostic process. First, we define logical structures which provide the interpretation for the strategies.

Definition 4.3 Lstrat~Model

A model for Cstrat is a structure M — (W, D, R_1, R_2, F), where W is a set of individuals (called worlds), D is a domain of individuals, R_1 and R_2 are accessibility relations on the worlds, i.e. subsets of W x W and F is an interpretation function.

F provides the interpretation for predicates and ground terms (in our case all ground terms are constants). The values of the variables are given by an assignment:

Definition 4.4 *Assignment*
Given a domain D an Assignment *is a mapping from the set of variables into D. By $\alpha_{(x|d)}$ we denote the assignment that maps variable x to an element $d \in D$ and is defined in the same way as α on the other variables.*

The semantics of an $\mathcal{L}_{Strat}$-Model is defined as follows:

Definition 4.5 *Semantics of $\mathcal{L}_{Strat}$*
Let $M = (W, D, R_1, R_2, F)$ be an $\mathcal{L}_{Strat}$-model, S, S_1, S_2 $\mathcal{L}_{Strat}$-Formulas, L an $\mathcal{L}$-Formula, α an assignment and $w \in W$. Let

$$Val(t, \alpha, w) := \begin{cases} \alpha(t), & \text{iff } t \text{ is a variable} \\ F(w, t), & \text{iff } t \text{ is a constant} \end{cases}$$

Then

$M \models_{w,\alpha} P(t_1, \ldots t_n)$ *iff* $(Val(t_1, \alpha, w), \ldots Val(t_n, \alpha, w)) \in F(w, P)$

$M \models_{w,\alpha} t_1 = t_2$ *iff* $Val(t_1, \alpha, w) = Val(t_2, \alpha, w)$

$M \models_{w,\alpha} S_1 \wedge S_2$ *iff* $M \models_{w,\alpha} S_1$ *and* $M \models_{w,\alpha} S_2$

$M \models_{w,\alpha} \neg S$ *iff* $M \not\models_{w,\alpha} S$

$M \models_{w,\alpha} \blacksquare S$ *iff for all* $w_1 \in W$ *s.th.* wR_1w_1: $M \models_{w_1,\alpha} S$

$M \models_{w,\alpha} \blacklozenge S$ *iff at least one* $w_1 \in W$ *exists s.th.* wR_1w_1 *and* $M \models_{w_1,\alpha} S$

$M \models_{w,\alpha} \Box L$ *iff for all* $w_1 \in W$ *s.th.* wR_2w_1: $M \models_{w_1,\alpha} L$

$M \models_{w,\alpha} \Diamond L$ *iff at least one* $w_1 \in W$ *exists such that* wR_2w_1 *and* $M \models_{w_1,\alpha} L$

$M \models_{w,\alpha} \forall x.S$ *iff for all* $d \in D$: $M \models_{w,\alpha_{(x|d)}} S$

We will use the following abbreviations:

$M \models_w S$ *iff* $M \models_{w,\alpha} S$ *for every assignment* α.

$M \models S$ *iff* $M \models_w S$ *for every* $w \in W$

$M \models STRAT$ *iff* $M \models S$ *for every* $S \in STRAT$

The connection between the state transition relation and the $\mathcal{L}_{Strat}$-model is established in the following way: The possible diagnoses are interpreted as possible worlds, where all the diagnoses under one set of working hypotheses are connected wrt. the $\Box$-operator (relation R_2). The accessibility relation for the $\blacksquare$-Operator (relation R_1) is given by the state transition relation $\mathcal{R}$, i.e. diagnoses under different working hypotheses are connected by R_1 iff the underlying sets of working hypotheses are connected by $\mathcal{R}$.

Definition 4.6 *Induced $\mathcal{L}_{Strat}$-Model $M_{\mathcal{R}}$*
Let $\mathcal{R}$ be a state transition relation. For $WH \subseteq WHYP$ let $\{M_{WH,1}, \ldots M_{WH,m_{WH}}\}$ be the set of models obtained from the system description, the observations and the diagnoses in $diag_{WH}(SD \cup OBS)$, where m_{WH} is the number of diagnoses. If $diag_{WH}(SD \cup OBS) = \emptyset$, then $m_{WH} = 0$. The Induced $\mathcal{L}_{Strat}$-Model $M_{\mathcal{R}}$ *is defined as* $M_{\mathcal{R}} = (W_{\mathcal{R}}, D_{\mathcal{R}}, R^1_{\mathcal{R}}, R^2_{\mathcal{R}}, F_{\mathcal{R}})$, *where*

$W_{\mathcal{R}} = \{(WH, i) \mid WH \subseteq WHYP, i \in \{1, \ldots m_{WH}\}, m_{WH} \geq 1\}$
$\cup \{\perp\}$, *i.e. for every diagnosis under a set of working hypotheses $WH \subseteq WHYP$ there is a world. There is one world $\perp$ representing the inconsistent states (an inconsistent state is a state, where $diag_{WH}(SD \cup OBS) = \emptyset$).*

$D_{\mathcal{R}} = CONST(\mathcal{L})$, *the set of constants of the language $\mathcal{L}$.*

$R^1_{\mathcal{R}} = \{((WH,i),(WH',j)) | WH\mathcal{R}WH' \wedge m_{WH'} \geq 1\}$
$\cup \{((WH,i),\perp) | \exists WH' : WH\mathcal{R}WH' \wedge m_{WH'} = 0\}$
$\cup \{(\perp,\perp)\}$
That means the accessibility relation on the diagnoses under different sets of working hypotheses is given by the state transition relation $\mathcal{R}$. Note, that the inconsistent world is a dead end, i.e. no other world is accessible from $\perp$

$R^2_{\mathcal{R}} = \{((WH,i),(WH,j)) \mid WH \subseteq WHYP, i,j \in \{1 \ldots m_{WH}\}\}$
$\cup \{(\perp,\perp)\}$, *that is, all diagnoses under the same set of working hypotheses are connected.*

$F_{\mathcal{R}}$: *Interpretation of predicate P of arity n:*

$$F_{\mathcal{R}}(\perp, P) = {D_{\mathcal{R}}}^n$$
$$F_{\mathcal{R}}((WH,j),P) = \{\vec{x} \in {D_{\mathcal{R}}}^n | P(\vec{x}) \in M_{WH,j}\}$$

Interpretation of $a \in CONST(\mathcal{L})$:

$$F_{\mathcal{R}}((WH,i),a) = a$$

$M_{\mathcal{R}} \models_{WH} S$ *iff for all* $i \in \{1, \ldots m_{WH}\}$: $M_{\mathcal{R}} \models_{(WH,i)} S$

The semantics just defined is a correct formalization of the concepts we introduced in section 3 as is summarized by the following proposition.

Proposition 4.7 *Properties of the semantics*
Let $L \in \mathcal{L}$ be a formula, $wh \in WHYP$ a working hypothesis, $\mathcal{R}$ a state transition relation and $WH \subseteq WHYP$ a set of working hypotheses. Then

$M_{\mathcal{R}} \models_{WH} \Box L$, *iff* L *is true under all diagnoses in* $diag_{WH}(SD \cup OBS)$
$M_{\mathcal{R}} \models_{WH} \Diamond L$, *iff* L *is true in at least one diagnosis in* $diag_{WH}(SD \cup OBS)$
$M_{\mathcal{R}} \models_{WH} \blacksquare\Box wh$, *iff For all* $WH' \subseteq WHYP$ *with* $WH\mathcal{R}WH'$: wh *is true under all diagnoses in* $diag_{WH'}(SD \cup OBS)$
$M_{\mathcal{R}} \models_{WH} \blacklozenge\Box wh$, *iff There is at least one* $WH' \subseteq WHYP$, *s. th.* $WH\mathcal{R}WH'$ *and* wh *is true under all diagnoses in* $diag_{WH'}(SD \cup OBS)$.

If a working hypothesis wh is known to be derivable under any diagnosis, then we want to conclude that wh has to be part of the current state. Formally, we want to garantee that $M_{\rightarrow} \models_{WH} \Box wh$ implies $wh \in WH$.

Problems occur if working hypotheses influence each other on the level of the system description. Assuming the system description contains a rule $wh' \rightarrow wh$, the strategy formula $\Box wh$ is satisfied if either wh' or wh is part of the current state. This is a disadvantage as the different levels of the system description and the strategies are mixed. The strategy formula $\Box wh' \rightarrow \Box wh$ expresses the same on the strategy level as $wh' \rightarrow wh$ does on the levle of the system description. The advantage of the strategy formula is that the formula $\Box wh \wedge (\Box wh' \rightarrow \Box wh)$ can only be satisfied if wh is part of the current state. In the remainder we assume that working hypotheses do not interfere.

Definition 4.8 *No interference*
Let $\rightarrow$ be a transition relation. Working hypotheses do not interfere each other, if

$$M_{\rightarrow} \models_{WH} \Box wh \text{ iff } wh \in WH$$

To conclude the definition of the semantics we answer the first question posed at the beginning of this section. We can now characterize the diagnostic processes: A diagnostic process is consistent with the underlying strategies if its transition relation satisfies the semantics.

Definition 4.9 *Consistent Transition Relation*
Let $\langle SD, STRAT, COMPS, OBS\rangle$ be the description of a diagnosis problem. Let the transition relation $\mathcal{R}$ be the encoding of the diagnostic process. $\mathcal{R}$ is a Consistent Transition Relation *for the given problem, iff* $M_{\mathcal{R}} \models STRAT$.

Example 4.10 *Consistent State Transition Relation*
The state transition relation $\mathcal{R}$ in example 4.2 is consistent wrt. the given problem. Consider for example the transition from the state $\{measure(Y)\}$ to $\{measure(Y), refine(C_2)\}$: In the strategy for structural refinement we say that a component should be refined, if it is known to be abnormal:

$$\forall c : (\Box refinable(c) \wedge \Box ab(c) \rightarrow \blacksquare\Box refine(c))$$

Since C_2 is refinable and known to be abnormal, this strategy formula postulates that it is refined in all states reachable from the current state. This strategy formula is satisfied by $\mathcal{R}$, because the only transition we consider is to the state $\{measure(Y), refine(C_2)\}$. The measurement strategy

$$\Diamond val(Y,1) \wedge \Diamond val(Y,2) \rightarrow \blacklozenge\Box measure(Y)$$

is also satisfied by $\mathcal{R}$ because we know the value of Y and consequently the left side of this rule is not true.

4.2 Results of the Diagnostic Process

So far we characterized which state transitions are allowed during the diagnostic process. In example 4.2 the diagnostic process terminated after a unique diagnosis was identified. In general this would be a too restrictive criterion for terminating the diagnostic process because we might not have enough knowledge to discriminate among all the diagnoses. Therefore we define, that the diagnostic process terminates in a state where we already assume all the hypotheses supported by the strategy formulas, because in that situation we cannot reach a more preferred state by applying another strategy.

Definition 4.11 *Stable State*
Let $WH \subseteq WHYP$ be a set of working hypotheses, $\mathcal{R}$ a consistent state transition relation and $STRAT$ a set of strategy formulas. The state characterized by WH ($diag_{WH}(SD \cup OBS)$) is a stable state *wrt. $\mathcal{R}$, iff*

1. $diag_{WH}(SD, OBS) \neq \emptyset$
2. $WH = \{wh \mid M_{\mathcal{R}} \models_{WH} \blacklozenge\Box wh\}$

Note, that the underlying transition relation is consistent. The first condition states that $SD \cup OBS \cup WH$ is consistent and the second condition is a fixpoint condition: WH is already the set of all working hypotheses suitable for the state described by WH. For complexity reasons we want the set WH to be as small as possible.

This will be discussed in section 4.4. The result of the diagnostic process is given by the diagnoses corresponding to the stable states:

Definition 4.12 *Result of the Diagnostic Process*
Let $(SD, STRAT, OBS, COMP)$ be a diagnosis problem. Let $\mathcal{R}$ be a consistent state transition relation for this problem. Then

$$\bigcup_{WH \text{ is a stable state wrt. } \mathcal{R}} diag_{WH}(SD \cup OBS)$$

is the result of the diagnostic process described by $\mathcal{R}$.

Now we are going to show that for a big subclass of the strategies that is suitable for specifying all the strategies presented in this paper, we can guarantee that the diagnostic process will reach a stable state if the system description is not contradictory in itself.

4.3 Monotonicity

By monotonicity, we mean that on each transition in the diagnostic process we only add some new working hypotheses to the set of hypotheses we already assume.

Definition 4.13 *Monotonicity of a State Transition Relation*
A state transition relation 7v *is* monotonic, *iff WHRWH' implies WH C WW*

We want the effect of strategies to be persistent, unless this leads to inconsistency. When applying structural refinement, we do not want to switch back to the abstract model we already used. When applying a focusing assumption we want to keep this assumption until we have either found a diagnosis, or we know that this assumption leads to inconsistency. Even actions like measurements can be described in a monotonic way by a hypothesis which has the effect that we know the measured value of a component *c*. Our meta-language is powerful enough to express the monotonicity of a set of strategies by adding additional formulas.

Definition 4.14 *Monotonic Extension of a Set of Strategies*
Let $STRAT$ be a set of strategies. The monotonic extension $Mon(STRAT)$ of $STRAT$ is defined as follows:

$$Mon(STRAT) := STRAT \cup \bigcup_{wh \in WHYP} \{\Box wh \rightarrow \blacksquare\Box wh\}$$

Lemma 4.15 *Let $STRAT$ be a set of strategy formulas, $\mathcal{R}$ a state transition relation. Then $M_{\mathcal{R}} \models Mon(STRAT)$ implies $\mathcal{R}$ monotonic*

Proof: (by contradiction) Suppose we have $M_{\mathcal{R}} \models Mon(STRAT)$, but for some WH, WH' we have $WH\mathcal{R}WH'$ and WH' is no superset of WH, i.e. $\exists wh : wh \in WH \backslash WH'$. As $wh \in WH$, we have $M_{\mathcal{R}} \models_{WH} \Box wh$. As $Mon(STRAT)$ is satisfied we conclude that $M_{\mathcal{R}} \models_{WH} \blacksquare\Box wh$ and as $WH\mathcal{R}WH'$ we know $M_{\mathcal{R}} \models_{WH'} \Box wh$. Because working hypothesis do not interfere each other we have $wh \in WH'$. This is a contradiction to the choice of wh.

For monotonic strategies we have the property that every path of the diagnostic process terminates, i.e. it leads to a stable or inconsistent state:

Theorem 4.16 *Termination of the Diagnostic Process*
Let $STRAT$ be a set of strategy formulas and $\mathcal{R}$ a transition relation, such that $M_{\mathcal{R}} \models Mon(STRAT)$. Then every transition sequence guided by $\mathcal{R}$ leads into a stable or inconsistent state.

Proof: $WHYP$ is finite and $\mathcal{R}$ is monotonic. So if we only add hypotheses in each step we will reach a fixpoint ($WHYP$ in the worst case).

4.4 Minimality

The diagnostic process shall be guided by the strategies. Only the hypotheses necessary to satisfy the strategies shall be assumed in each step. This is captured by the concept of *local minimality*. We first define a notation for the working hypotheses entailed by the strategies wrt. our declarative semantics.

Definition 4.17 *Supported*
Let $\mathcal{R}$ be a state transition relation and $WH \subseteq WHYP$.

$$supported(WH, \mathcal{R}) := \{wh \mid M_{\mathcal{R}} \models_{WH} \blacklozenge\Box wh\}$$

The set $supported(WH, \mathcal{R})$ denotes the hypotheses assumed in at least one of the successor states.

Proposition 4.18 *$wh \in supported(WH, \mathcal{R})$, iff for at least one $WH' : WH\mathcal{R}WH' \wedge wh \in WH'$.*

Definition 4.19 *Locally Minimal*
A state transition relation $\mathcal{R}$ is locally minimal, *iff*

1. *$\mathcal{R}$ is consistent.*
2. *For every set of working hypotheses WH, $supported(WH, \mathcal{R})$ is a minimal set of hypotheses, i.e. there is no consistent $\mathcal{R}'$, such that $supported(WH, \mathcal{R}') \subset supported(WH, \mathcal{R})$.*

5 Examples of Strategies

First we give the formalization of some additional strategies introduced by Boettcher and Dressler [2].

5.1 Behavioral Refinement

Behavioral Refinement is useful if we have behavioral models at different levels of detail for a component C. For complexity reasons we initially use the simplest model available. But when different diagnoses predict different behavioral modes for C, we activate a more detailed model for C's behavior in order to discriminate between these diagnoses. This can be expressed by

$$\forall c : (\exists m_1 : \exists m_2 : (\Diamond mode(m_1, c) \wedge \Diamond mode(m_2, c))) \rightarrow \blacklozenge\Box ref\_fm(c)$$

In the system description, $ref\_fm(C)$ activates the detailed behavioral model for a component C.

5.2 Physical Negation

If a behavioral model other than the unknown mode can be assigned to a component C, we do not consider the unknown mode (i.e. we assume the specified fault models are complete).

$$\forall c : (\exists m : (\Diamond mode(m, c) \wedge m \neq unknown)) \rightarrow \blacklozenge\Box fm\_complete(c)$$

If *fm-complete* is active for a component C, we can assure that C is assigned a known fault mode by adding the following rule to the system description:

$$\forall c \; : \; fm\_complete(c) \rightarrow (ok(c) \vee \exists m : (mode(m, c) \wedge m \neq unknown))$$

Our approach allows to formalize all other strategies presented by Boettcher and Dressier as well. However in some cases we find it more appropriate to express them by preferences on diagnoses (see [6] for a discussion of this issue). The next strategies are completely new ones.

5.3 Integrating Heuristic Knowledge into Model—Based Diagnosis

Consider an electronic device, where a single chip contains a number of gates (e.g. and-gates). Assume we have *n* such chips. From experience we know that a diagnosis containing two and-gates on different chips is much less likely than a diagnosis containing two and-gates on the same chip, as the latter can be explained by a single cause that damaged the whole chip. Such heuristic knowledge if. easy to describe in our strategy language. We use *loeation{A, C)* to denote that component *A* is located on chip *C*. The heuristic assumption expressing our belief that all faulty and-gates are on one chip is represented by the working hypothesis *focus.chip.* The effect of this assumption is expressed in the system description as follows:

$$focus\_chip \rightarrow \exists l : \forall a : ((type(a, And\_gate) \wedge ab(a)) \rightarrow location(a, l))$$

The following strategy specifies that the hypothesis $focus\_chip$ should be assumed if consistent:

$$(\exists l : \forall c : \Diamond((type(c, And\_gate) \wedge ab(c)) \rightarrow location(a, l))) \rightarrow \blacklozenge\Box focus\_chip$$

Using $\blacklozenge\Box focus\_chip$ instead of $\blacksquare\Box focus\_chip$ allows us to explore states not containing $\Box focus\_chip$.

5.4 Measurements and Dependent Actions

If different consistent models of the system predict, different values for some measure point -V we can discriminate between these values by making a measurement, as represented by the following strategy rule

$$\forall x : \Diamond val(x, 1) \wedge \Diamond val(x, 0) \rightarrow \blacklozenge\Box measure(x)$$

Since measurements require interaction with the user, we want to express that measurements should only be made when no other strategies are available.

In some situations, more than one assumption is supported at the same time. I *strat* allows to explicitly specify that a working hypotheses *wh* should only be considered, if *wh'* is not supported. For example the following formulas specify the strategies structural refinement and measurements so that measurements are only performed, if all useful structural refinements have already been considered.

$$\begin{aligned}
&\forall c : (\Box refinable(c) \wedge \Box ab(c) \rightarrow \blacksquare\Box refine(c)) \\
&\forall c : (\Box refinable(c) \wedge \Diamond\neg ab(c) \rightarrow \neg\blacklozenge\Box refine(c)) \\
&\forall x : (\Diamond val(x, 0) \wedge \Diamond val(x, 1)) \\
&\quad \wedge (\forall c : \blacklozenge\Box refine(c) \rightarrow \Box refine(c)) \\
&\quad \rightarrow \blacksquare\Box measure(x)
\end{aligned}$$

Since we only consider minimal models of these formulas, *measure(x)* will only be assumed if all the conditions on the left side of the last formula are satisfied.

6 Conclusion and Further Work

This paper defines the concept of diagnosis strategies using a modal logic language that makes strategic knowledge explicit. Our approach allows not only to express system models in a declarative way (which is one of the main advantages of model-based diagnosis), but extends this declarativity to the meta level by allowing the declarative description of diagnosis strategies.

We are currently working on an efficient implementation of the formal concepts introduced in this paper using transformations of our meta-language into first order logic and minimal model semantics within our DRUM diagnosis system.

Finally, we want to thank Carlos Damasio and Luis Percira for their fruitful cooperation within the INIDA project on the topics discussed in this paper and in a companion paper ([3]).

References

[I] C. Bottcfier and O. Dressier. Diagnosis process dynamics: Holding the diagnostic trackhound in leash. In *Proceedings of the International Joint Conference on Artificial Intelligence,* volume 2, pages 1460-1471. Morgan Kaufmann Publishers, Inc., 1993.

[2] C Bottcher and O. Dressier. A framework for controlling model-based diagnosis systems with multiple actions. *Annals of Mathematics and Artificial Intelligence, special Issue on Model-based Diagnosis,* 11(1-4), 1994.

[3] C. V. Damasio, W. Nejdl, L. Pereira, and M. Schroeder. Model-based diagnosis preferences and strategies representation with meta logic programming. In K. R. Apt and F. Turini, editors, *Metalogics and Logic Programming,* chapter 11, pages 269-311. The MIT Press, 1995.

[4] *C* V. Damasio, L. M. Pereira, and W. Nejdl. Revise: An extended logic programming system for revising knowledge bases. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning,* pages 607-618, Bonn, Germany, May 1994. Morgan Kaufmann Publishers, Inc.

[5] J. de Kleer and B. C. Williams. Diagnosing multiple faults. *Artificial Intelligence,* 32:97-130, 1987.

[6] P. Frohlich, W. Nejdl, and M. Schroder. A formal semantics for preferences and strategies in model-based diagnosis. In *5th International Workshop on Principles of Diagnosis (DX-94),* pages 106-113, New Paltz, NY, Oct. 1994.

[7] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence,* 32:57-95, 1987.

[8] P. Struss. Diagnosis as a process. In W. Hamscher, L. Console, and J. de Kleer, editors, *Readings in Model'Based Diagnosis,* pages 408-418. Morgan Kaufmann Publishers, Inc., 1992.