

Diagnosis Process Dynamics:

Holding the Diagnostic Trackhound in Leash

Claudia Böttcher
Fraunhofer-Institute IITB
Karlsruhe, Germany
boet@iitb.fhg.de

Oskar Dressler
Siemens AG
München, Germany
dressler@zfe.siemens.de

Abstract

A main obstacle for building large diagnosis systems is the problem of deciding *when* to use *which* inference pattern or representation. If diagnostic actions such as changing representations or applying specific inference patterns are understood in terms of change of working hypotheses, the control problem becomes a belief revision problem: when to adopt or drop beliefs. Our approach proceeds in two steps. First, we adopt the principle of informational economy as kind of a law of inertia for diagnostic processes. It proposes candidates for revised belief states. In a second step we employ specific *diagnostic knowledge* to actually choose the next belief state.

1. Introduction

In recent years reasoning about structure and function of physical systems for the purpose of diagnosis has seen a dramatic increase in activities. A considerable set of new exciting results have emerged from these:

- modelling issues such as representation and use of temporal aspects [Hamscher 91], models of different granularity, relations between models [Struss 92], and multiple component modes [de Kleer, Williams 89]
- diagnostic inference patterns, e.g. alibis [Raiman 89], physical negation [Struss, Dressler 89] and physical impossibility [Friedrich et al. 90], integration of abduction e.g. [Console, Torasso 90], use of non-intermittency [Raiman et al. 91], and
- inferential power [de Kleer 91].

Nevertheless, a main obstacle for building large diagnostic systems remains. How can we control *when* to use *which* inference pattern or representation?

Consider the example from [Davis 82]. A computer is apparently working fine when viewed from the console terminal. The hardware powers up, the operating system loads and runs. The user terminals appear functional, but nothing happens when we type something.

Supplied with a very coarse model, a diagnosis system like GDE [de Kleer, Williams 87] can identify the computer as the most probable suspect. At this point the best diagnostic action is a structural decomposition. *But why this one instead of a switch in behavioral model granularity for one of the terminals for example?*

Structural decomposition moves us to another world illustrated in fig. 1. If disk, controller, bus#1, memory, memory control or cpu were malfunctioning, the operating system would not be running. Supplied with the diagnostic hypotheses that fault modes for these components are complete, GDE⁺

[Struss, Dressler 89] can exonerate all of them. *But why choose the diagnostic action of adopting fault mode completeness for exactly these components instead of e.g. a structural decomposition of a non-suspect component?*

In only two diagnostic steps the available diagnosis machinery could pin down two highly suspicious components, namely the bus interface INT and Bus#2, *because we have chosen the right diagnostic actions by hand*. A few more steps lead to a bent spring-loaded connector between a board and a backplane slot. Consider the enormous reduction in complexity! The key to controlling inference in large diagnostic problems is the ability to choose between diagnostic actions and *know when to do it*.

Carefully choosing the right technique at the right time would allow minimizing the overall cost of model simulation, measurement selection, probing etc. But the conventional diagnosis loop of predicting behavior by simulating (component) models, detecting discrepancies, e.g. between observed and predicted values, generating diagnoses and, finally, discriminating among the most plausible diagnoses by additional measurements is inadequate for at least two reasons:

- In general, there is a lot more reasoning that can be done before the agent needs to go probing. It can e.g. switch to structurally or behaviorally more refined models or adopt or drop diagnostic hypotheses such as fault model completeness, preference for explaining diagnoses or about the nature of faults, e.g. non-intermittency.
- Even though we have seen a fundamental breakthrough in inferential power [de Kleer 91] the agent cannot afford to run all component models etc.

Therefore, we need a framework for deciding *when* to take *which* action, i.e. when to switch a model, take a measurement, drop or adopt a diagnostic hypothesis, etc.

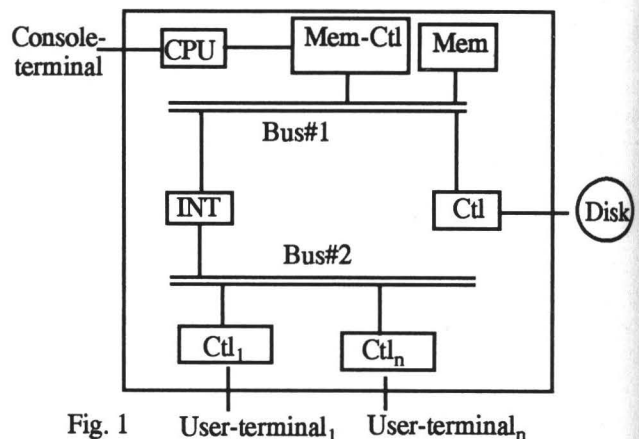


Fig. 1 User-terminal₁ User-terminal_n

Approach and Outline

Changing hypotheses means changing the agent's belief state. Consequently we treat the control problem as a belief revision problem. A belief state change is triggered by either the current state not allowing for the generation of plausible diagnoses or the existence of a more preferred "neighbor". The guiding principle for revision is *informational economy* [Gärdenfors 88] which we independently motivate for diagnostic contexts. Acting as a kind of law of inertia for diagnostic processes it requests that changes to the current working hypotheses have to be minimal and provides the basic candidates for revision. A belief revision operator defined in terms of these candidates automatically eliminates useless changes due to logical relations e.g. between component models or component modes. From the remaining changes the most preferred is chosen drawing on general knowledge about assessing diagnostic belief states qualitatively. Since every diagnostic action can be identified as a change of working hypotheses selecting the best next belief state amounts to selecting the best next diagnostic action.

Section 2 shows how we represent mode and model assumptions in relation to models. Introducing the notion of preferred diagnosis we specify our representation of an agent's belief state. In section 3 we turn to the dynamics of *diagnostic* belief state changes and identify two forms of non-monotonicity involved. Section 4 and 5 provide the technical details concerning informational economy, our belief revision operator, computation of basic candidates for revisions and selecting the best next belief state for the diagnosis process.

2. Diagnostic Belief States

A model-based diagnostic agent is supplied with a system description SD and a set of observations Obs . Often we refer to $SD \cup Obs$ as the background theory T . Two basic insights help structure SD . Consider the model of the computer in the example. First, at the beginning we only need a very coarse model using qualitative input/output values like 'looks fine' or 'looks bad'. In later stages of the diagnosis process more detailed models may be required. Second, the computer might work in different good and fault modes that need to be represented as different models. Combining both requirements we represent model chunks for component c as formulae of the form

$$modg \wedge mode(c) \wedge \dots \rightarrow \dots$$

where $mode(c)$ identifies the component mode and $modg$ the model granularity. For instance, the implication

$$fine/bad\text{-}level \wedge ok(computer) \wedge input=looks\text{-}fine \\ \rightarrow output=looks\text{-}fine$$

describes the computer at a coarse level for its ok mode.

A focused ATMS ([Dressler-Farquhar 90], [de Kleer 91]) allows using the model granularity and mode preconditions for control purposes. Predictions will only be made when the preconditions are activated during the diagnosis process. To control the activation the agent maintains a set of working hypotheses w . This may include model granularity and mode assumptions as well as diagnostic hypotheses, e.g. about the nature of faults. These "precondition" assumptions and logical relations between them are known to the diagnostic agent as part of SD but no predictions are made unless they are selected. Once made, however, predictions are cached in the ATMS.

Assigning modes of behavior to the components of a device is a central task of the agent. Mode assignments must be consis-

tent with the observations and the working hypotheses currently held. Working hypotheses are taken from the set WHP containing among others model granularity assumptions and diagnostic hypotheses. Subsets of WHP are called *worlds*. We represent worlds w by their "w-closure", i.e. $WHP \cap Cn(w \cup T)$, where $Cn(A) := \{\alpha \mid A \text{ derives } \alpha\}$.

Def. 2.1 : Let modes (C) denote the modes of a component C . A mode assignment $\Pi \in \text{modes}(C_1) \times \dots \times \text{modes}(C_n)$ (denoted $\{.. m_j(C_i)..\}$) is a **diagnosis** under hypotheses w iff $SD \cup Obs \cup w \cup \Pi$ is satisfiable.

Not all diagnoses are equally plausible. One reason is that different mode assignments to an individual component may have different plausibility, e.g. ok modes are preferred over fault modes. We assume that a partial order $\geq$ for the modes of a component is given. If needed, this order may depend on a corresponding working hypothesis and thus may be subject to later revisions. Preference relations among modes induce a preference relation among diagnoses: let $\Pi = \{.. m_j(C_i)..\}$ and $\Pi' = \{.. m_k(C_i)..\}$ be two diagnoses in world w . Then $\Pi \geq \Pi'$ iff $\forall i=1, \dots, n: m_j(C_i) \geq m'_k(C_i)$.

Def. 2.2 : A diagnosis Π is a **preferred diagnosis** under working hypotheses w iff no diagnosis under w is strictly preferred over it: $\forall \Pi': \Pi' \geq \Pi \Rightarrow \Pi' = \Pi$

Example: Let the user terminals have three modes, $ok > fm_1 > fm_2$, and the computer have two modes, $ok > fm$. If all combinations of fault models are consistent with the observations, then $\{ok(console), fm(computer), ok(term\#1), \dots, ok(term\#n)\}$ and $\{ok(console), ok(computer), fm_1(term\#1), \dots, fm_1(term\#n)\}$ are preferred diagnoses, but $\{ok(console), ok(computer), fm_2(term\#1), \dots, fm_1(term\#n)\}$ is not.

In what follows we omit ok modes, e.g. $\{fm(computer)\}$ denotes the first preferred diagnosis from above. We encode component modes as normal defaults with prerequisites [Reiter 80]. $\dots \wedge \neg m_i(C) \wedge \dots : m_j(C) / m_j(C)$ denotes the default for the j -th mode of component C . This guarantees that a mode $m_j(C)$ will be considered only when the more preferred modes $..m_i(C)..\$ have been refuted. These defaults can be encoded in a non-monotonic ATMS. This NM-ATMS computes minimal consistent extension bases, $MinExtBases(D, w \cup T)$, that are in one-to-one correspondence with the extensions of the default theory $\Delta=(D, w \cup T)$ and hence preferred diagnoses.

Theorem 2.1 : Π is a preferred diagnosis under hypotheses w iff $\exists B \in MinExtBases(D, w \cup T): \Pi \subseteq Cn(w \cup B \cup T)$.

The belief state comprises predictions made from $SD \cup Obs$. They are indexed by the underlying mode and model granularity assumptions. If $w \cup T$ has no consistent extension, then w is said to be *incoherent*. It is the incomplete set of observations about the artifact and the inconsistencies discovered that give rise to multiple extensions, and hence multiple plausible diagnoses. See [Dressler, Struss 92] for details.

3. Diagnostic Belief State Change

Diagnosis is a sequential process. During this process the set of plausible candidates for diagnosis may expand as well as shrink depending on choices taken by the diagnostic agent. Each choice plays the role of a working hypothesis that may be retracted later on, only to be re-adopted under some future circumstances. In the first step of the example there are two preferred diagnoses, namely $\{fm(computer)\}$ and $\{fm(ter-$

$m\#1), \dots, fm(term\#n)\}$. We zoom into the computer's structure. This leads to preferred diagnoses such as $\{fm(bus\#2)\}$, i.e. we *expand* the set of preferred diagnoses. Imagine that further observations lead us to conclude that everything is ok with the computer. All of the new diagnoses are refuted. Then we *shrink* the set of preferred diagnoses.

There are two forms of non-monotonicity involved. *Logical non-monotonicity* occurs when an agent generates expectations about plausible inferences based on its current *incomplete* belief state. *Temporal non-monotonicity* occurs when an agent changes its belief state by dropping some beliefs while possibly assimilating others. It is crucial to understand that belief state changes are often motivated by reasons beyond logic. For example, an agent may want to change its belief state not only because it has become inconsistent but because it simply prefers another, e.g. when the number of preferred diagnoses differs by an order of magnitude.

Both, temporal and logical non-monotonicity are needed for a natural formalization of the behavior of diagnostic agents. The rest of the paper deals with temporal non-monotonicity. We discuss two scenarios separately. They correspond to the two reasons for triggering a belief state change:

- the selected set of working hypotheses does not allow for the generation of diagnoses
- there is a more preferred "neighbor" world

4. Revising Incoherent Belief States

4.1 Plausible Candidates for the Next Belief State

We start with the former scenario, i.e. $\Delta = (D, SD \cup Obs \cup w)$ has no extension. We consider the potential move from a set of working hypotheses w_1 to a set w_2 . In a different terminology we could say that all propositions but the working hypotheses are protected formulas. Some of the selected working hypotheses are going out, some others are about to become in. The union of these tells us *how much* is changed in terms of working hypotheses when moving from w_1 to w_2 . Since we represent worlds by their deductive closure, the definition is less syntactical than it appears. Logical connections between working hypotheses are reflected, e.g. the "at most one fault" assumption implies "at most two faults" etc. Furthermore, we include integrity constraints on worlds requiring for instance that (at least) one model granularity assumption is activated for every component.

Def. 4.1 : $\text{diff}(w_1, w_2) := (w_1 \setminus w_2) \cup (w_2 \setminus w_1)$

When a selected world is *incoherent*, i.e. has no consistent extension, the diagnostic agent has to decide whether there is another world allowing for the generation of diagnoses. It has to make a choice between the possible moves, but should not select an arbitrary one. There are at least two reasons: First, the primary sources of diagnostic information are the sets of assumptions underlying a contradiction (conflicts). At least one of the components is not working in the specified mode or a working hypothesis is violated. Therefore, on the one hand we want to keep the diagnostic agent focused on these assumption sets. On the other hand there is the need to change the working hypotheses. Allowing only minimal changes of the selected set of working hypotheses minimizes the risk of losing conflicts. This is the main argument for the principle of *informational economy* in diagnostic contexts. There is a second, computational one. Allowing non-minimal changes

might shift already cached intermediate results out of focus. This is especially the case when using a focused ATMS. Therefore, exploring totally new foci is more expensive than exploring partially new ones. Possible moves are collected in the set 'revisions' below. The definition is given in terms of preferred diagnoses and minimal set difference of source and target world.

Def. 4.2 : Let w be a world. **Revisions** (w) :=

$\{v \in \text{worlds} \mid \exists \text{ preferred diagnosis under } v \text{ and } \text{diff}(w, v) \text{ is minimal wrt. set inclusion}\}$

Whether a diagnostic action is useful only becomes clear when behavior prediction based on the selected hypotheses is done and the agent reaches a new belief state after generating diagnoses. But simulating models for the sole purpose of deciding which action to choose already performs those computations most of which we want to save. Hence, a *limited* lookahead is needed. The effects of a diagnostic action are essentially determined by the set of selected hypotheses resulting from the action. A diagnostic action moves the agent from its current world w_1 to a target world w_2 . And of course, the target world should allow for the generation of diagnoses, i.e. be coherent. Since we cannot invoke behavior prediction for deciding whether w_2 is indeed coherent, we use an approximation. We only use the inferences cached in the ATMS for making the decision. This means that the decision whether a target world is coherent is made on the basis of already simulated models only. This adequately reflects our view that the agent uses its *limited* knowledge that characterizes the current diagnostic situation.

4.2 Revision Operator

The set $\text{revisions}(w)$ is but a set of plausible "neighboring" worlds. The diagnostic agent still has a choice as to which of these worlds it wants to enter. The function ϕ selects it.

Def. 4.3 : Let ϕ be a selection function and $w \in \text{worlds}$. The revision operator rev-op is $\text{rev-op}(w) := \phi(\text{revisions}(w))$.

4.2.1 Exploiting Logical Relations

The operator is surprisingly powerful in diagnostic contexts. By definition it avoids target worlds where no diagnoses exist. This eliminates almost all useless diagnostic actions the agent could perform otherwise. We consider two examples. The first one is structural decomposition. It shows how logical relations between component *modes* are automatically exploited by the revision process. The second is the use of simplifying assumptions which stands for the class of logical relations between component *models*.

Example: If the computer's subcomponents are ok, the computer should be working correctly. In order to provide a handle for control and to reflect implicit assumptions like the above, we include the model granularity assumption $\text{modg-c}_{\text{sub}}$ in the implication $\text{modg-c}_{\text{sub}} \wedge \text{ok(cpu)} \dots \wedge \text{ok(disk)} \rightarrow \text{ok(computer)}$

which is part of the system description SD. Structural decomposition of the computer model means adding $\text{modg-c}_{\text{sub}}$ to the current world w . If, however, no preferred diagnosis exists, then, so the following lemma tells us, no superset of w has one either. Hence, adding $\text{modg-c}_{\text{sub}}$ is not a choice taken by the revision operator in our example.

Lemma 4.1 : For all $w' \supseteq w$: if there is no preferred diagnosis under w then there is no preferred diagnosis under w' .

Example: A (component) model m' is a simplification of another model m , when values entailed by m' are also entailed by m and a simplifying assumption δ . [Struss 92]. Let M and M' be the associated model granularity assumptions. Then $M \wedge \delta \rightarrow M'$. Simplified models are often much easier to simulate than more detailed ones. Therefore an agent starts with the simplified model, i.e. $M' \in w$. In case no preferred diagnoses under w exist, due to the last lemma, activating the detailed model and the simplifying assumption, i.e. $\delta, M \in w$, does not lead to diagnoses since they imply M' . The revision operator will not do that. Deactivating δ , i.e. $\delta \notin w, M \in w$, could help if there are no other causes for the non-existence of diagnoses.

4.2.2 Preferences between Worlds

Exploiting logical relations avoids useless actions such as refining a non-suspect component. But there is more *diagnostic* knowledge available. The selection function ϕ need not be that blind when selecting an element from $\text{revisions}(w)$. If e.g. the number of preferred diagnoses in a world w_1 is an order of magnitude greater than in w_2 then prefer world w_2 . In [Böttcher 91] a variety of such examples are discussed.

Example: Physical negation concludes, that when all fault modes of a component are known and have been refuted, the component must be ok. E.g.

$$\neg \text{fm}(\text{cpu}) \wedge \text{fault-modes-complete}(\text{cpu}) \rightarrow \text{ok}(\text{cpu})$$

If $\text{SD} \cup \text{Obs} \cup w \cup \{\text{fm}(\text{cpu})\}$ is not satisfiable, then adopting $\text{fault-modes-complete}(\text{cpu})$ leads to $\text{ok}(\text{cpu})$. Hence all preferred diagnoses under $w' \supseteq w$ assigning a fault mode to the cpu are eliminated. Nevertheless, as long as the fault mode has not been refuted, the described effect will not take place. In this case the operator has no preference for adopting $\text{fault-modes-complete}(\text{cpu})$ because the set of preferred diagnoses remains the same.

The last example gives a flavor of the preferences that can guide the selection of worlds. It is also in order to make use of criteria like e.g. probabilities for the applicability of working hypotheses, when they are available. For the selection of diagnostic actions in general, however, we can only expect to have a dictionary of partial preferences $\text{DPP} = \{\text{DPP}_i \mid i \in 1..n, \text{DPP}_i \subseteq 2^{\text{WHP}} \times 2^{\text{WHP}}\}$. Ideally, one would want to aggregate a global preference order from DPP in order to allow for a rational choice by the selection function ϕ . This, however, is impossible given the results reported in [Doyle 91]. It is shown that no aggregation policy exists that could aggregate a collection of partial preferences, e.g. DPP , into a global order satisfying fairly mild restrictions. The order below therefore is defined with the goal that the diagnostic agent can use obvious preferences when available.

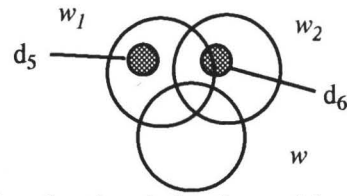
Def. 4.4 : (Aggregation Policy)

Let w, w_1, w_2 be worlds. $w_1 \geq_w w_2$ iff

$$\begin{aligned} & [(\exists k \exists (d_1, d_2) \in \text{DPP}_k : d_1 \subseteq w_1 \wedge d_2 \subseteq w_2) \\ & \wedge \neg (\exists j \neq k \exists (d_3, d_4) \in \text{DPP}_j : d_3 \subseteq w_2 \wedge d_4 \subseteq w_1)] \\ & \vee \exists i \exists (d_5, d_6) \in \text{DPP}_i : d_5 \subseteq w_1 \setminus w \wedge d_6 \subseteq w_2 \setminus w \end{aligned}$$

It prefers a state w_1 over w_2 when no opposite preferences are known (first disjunct). Nevertheless, large worlds tend to be ordered in multiple ways by different DPP_i . Therefore, w_1 and w_2 are likely *not* to be ordered by the first disjunct because their overlaps with the current world w may already contain two opposite preferences. A preference, however, may be justified because of the *new* hypotheses added in the candidate

worlds w_1 and w_2 . This is achieved by the second disjunct. Fig.2 illustrates the situation.



The selection function ϕ for the revision operator rev-op chooses a maximally preferred world in revisions (w) with respect to the aggregated order $\geq_w$.

We can encode a wide range of preferences. In a moment we shall list some common preferences that apply to large classes of diagnosis applications. Most of these preferences can be derived from the fact that we are doing diagnosis of an artifact that worked before and was not e.g. the target of a bomb attack. Therefore, one can assume that a small percentage of its components are defective. Hence, it makes sense to prefer actions that aim at proving correctness of suspected components, e.g. via physical negation, or to delay further elaboration on suspects, e.g. via structural refinement, until they are clearly identified. In general, preferences may depend on the current world w and its set of preferred diagnoses $\text{diag}(w)$.

Let COMPS denote the set of all components and $\text{mode}_{\Pi}(c)$ be the behavior mode assigned to component c by diagnosis Π

P1: Structural Refinement

Suppose our diagnostic agent currently operates on the i -th hierarchic layer of some component c and considers a structural refinement as a possible move. This is reflected by a working hypothesis for the current layer of c , say $h_i(c)$, in the current world and $h_{i+1}(c)$ in a coherent neighbor world. We do structural refinement only, if the component is identified uniquely as defective:

$$\begin{aligned} & \text{if } \forall \Pi \in \text{diag}(w) : \text{mode}_{\Pi}(c) \rightarrow \neg \text{ok}(c) \\ & \text{then } \{h_{i+1}(c)\} >_{P1} \{h_i(c)\} \text{ else } \{h_i(c)\} >_{P1} \{h_{i+1}(c)\} \end{aligned}$$

P2: Behavioral Refinement

If a model $m(c)$ (possibly under some working hypothesis) makes strictly stronger predictions than another model $m'(c)$, then refuting the weaker one also refutes the stronger ([Struss 92]). Therefore, when multiple models for the same behavioral mode are available, we prefer simpler models when possible. If we want to discriminate between diagnoses, however, we try to refute some of them. This can be done by making stronger predictions, e.g. switching to stronger models for the behavioral modes occurring in the diagnoses.

$$\begin{aligned} & \text{if } \exists \Pi_1, \Pi_2 \in \text{diag}(w) \exists c \in \text{COMPS} : \\ & \quad \text{mode}_{\Pi_1}(c) \neq \text{mode}_{\Pi_2}(c) \\ & \text{then } \{m(c)\} >_{P2} \{m'(c)\} \\ & \text{for all other components: } \{m'(c)\} >_{P2} \{m(c)\} \end{aligned}$$

P3: Physical Negation

We avoid assigning the unknown behavioral mode to a component with known fault modes by physical negation:

$$\begin{aligned} & \text{if } \exists \Pi \in \text{diag}(w) : \text{mode}_{\Pi}(c) = \text{unknown}(c) \wedge \\ & \quad \exists \text{mode}(c) \neq \text{unknown}(c) : \text{mode}(c) \rightarrow \neg \text{ok}(c) \\ & \text{then } \{\text{fault-modes-complete}(c)\} >_{P3} \{\} \\ & \text{else } \{\} >_{P3} \{\text{fault-modes-complete}(c)\} \end{aligned}$$

P4: N-Fault Assumption

Simple diagnoses in terms of the number of broken components are preferred:

$$\begin{aligned} & \text{if } i < j \leq \#\text{COMPS} \\ & \text{then } \{\text{at-most-}i\text{-faults}\} >_{P4} \{\text{at-most-}j\text{-faults}\} \end{aligned}$$

P5: Non-Intermittency

Presuming non-intermittency for a behavioral mode allows the agent to predict identical outputs of a component at time t_i given the same inputs as at some other time t_j ([Raiman et. al. 91]). As long as feasible, we assume non-intermittent modes: $\forall c \in \text{COMPS } \forall m \in \text{modes}(c): \{\text{non-intermittent}(m)\} >_{P5} \{\}$

P6: Further Measurements

When the current world is the most preferred one, the agent longs for more *external* information. It makes sense to take additional probes, if this discriminates between preferred diagnoses. Finding optimal discriminating points is a function of the measurement proposer as in [de Kleer, Williams 87]. If, however, no additional measurement can discriminate between the preferred diagnoses, then a measurement can at best refute them all. In this case the agent prefers to make any other available move instead of doing random probes.

if no-more-discriminating-measurements (w)

then $\forall a \in \text{WHP}: \{a\} >_{P6} \{\}$ **else** $\{\} >_{P6} \{a\}$

In expressing this preference we have implicitly assumed that probes are associated with some non-negligible costs. This is not the general case. In many applications the measurement proposer can act as an intelligent, situation dependent filter that selects from anyway available observations. Therefore, it may be advisable to do more probes, i.e. gather more evidence before investing all the effort to find plausible interpretations of an incompletely described situation:

if more-measurements (w)

then $\forall a \in \text{WHP}: \{\} >_{P6'} \{a\}$ **else** $\{a\} >_{P6'} \{\}$

All of these preferences can be further elaborated. Often they can be tuned by exploiting application characteristics, e.g. if a component is likely to fail in some unpredictable way, then we better avoid applying the physical negation rule. For another example consider P2, behavioral refinement. If we have tried a stronger model for a component upon which two different diagnoses disagree, but none of them is refuted, then we can make even stronger predictions by switching to more detailed models for more of the other components.

4.3 Computing Revisions

The key insight to computing revisions(w) is the following observation: given the working hypotheses $\text{diff}(w_1, w_2)$ that are either going *in* or going *out* when changing from w_1 to w_2 , and the current world w_1 , we can determine the target world w_2 :

$$\text{diff}(w_1, \text{diff}(w_1, w_2)) = w_2$$

This allows monotonically enumerating potential diff 's breadth first starting from the empty set. Since we are interested only in minimal diff 's that satisfy certain conditions, we avoid testing arbitrary target worlds.

compute-revisions(w)

```
vars: agenda := { {} }, coherent-diffs, d, diff
loop until empty(agenda) do d:= pop(agenda)
  loop for h in working-hypotheses do
    diff:= { h }  $\cup$  d ; target:= diff(w, diff);
    if subsumed (diff, coherent-diffs) then nothing else
    if coherent (target)
    then add-and-minimize (target, coherent-diffs)
    else push (diff, agenda)
```

return-values:

loop for cdiff in coherent-diffs collect $\text{diff}(w, \text{cdiff})$

We perform a breadth first search in the lattice of sets of working hypotheses, i.e. the search order for diff is $\{\}$,

$\{h_1\}, \dots, \{h_n\}, \{h_1, h_2\}, \dots$ If it leads to a coherent world, we update a set of minimal diff s, coherent- diff s. Otherwise we push diff on the agenda, i.e. we expand it the next time through the loop. We stop when it is subsumed.

Usually the algorithm stops after just two top level iterations since a multiple change of working hypotheses is seldom required. But the main effort for computing revisions(w) is hidden in the predicate *coherent*. It requires that MinExtBases be computed ([Dressler 90]). This, however, is no additional overhead. The NM-ATMS computes MinExtBases anyway for the generation of preferred diagnoses (theorem 2.1).

5. Revising Coherent Belief States

If the current world is incoherent a belief state change is inevitable. But there may be other reasons.

Example: If there is a preferred diagnosis assigning a fault to the computer, it may make sense to refine its structure since $\neg \text{ok}(\text{computer}) \wedge \text{modg-c}_{\text{sub}} \rightarrow \neg \text{ok}(\text{cpu}) \vee \dots \vee \neg \text{ok}(\text{disk})$

In this case the revision operator could select $\text{modg-c}_{\text{sub}}$ for inclusion in w . Because of suspecting the computer the agent expands the substructure thereby generating a multitude of preferred diagnoses. Now, if a technician measures the computer's output towards the terminals and finds that it is ok, then the computer is no longer suspect. Moving back to the black box level of the computer, i.e. dropping the assumption $\text{modg-c}_{\text{sub}}$ would result in significantly fewer preferred diagnoses. Therefore the agent *prefers* this world without being forced by logical reasons.

We still accept the principle of informational economy. Diagnostic actions should not jump to arbitrary target worlds. But in contrast to the situation discussed in section 4 the current world is coherent. There is at least one preferred diagnosis. The agent may take the opportunity to strengthen its position by adopting stronger working hypotheses that, when added, still allow for the generation of diagnoses. Consider for instance the initial situation when the diagnosis process has just been started. The empty world is coherent (modulo potentially violated integrity constraints) since no prediction was performed, hence no contradictions detected so far. The agent may now adopt working hypotheses. We allow it to iterate choosing "neighbor" worlds without interleaving prediction. This means that the agent can choose a maximally preferred world from all worlds that can be reached from the current world without traversing an incoherent one. Of course, it may also choose the current world because it is coherent. The selection process then is roughly similar to the one discussed in the last section. The main difference is that the current world may be a possible candidate for a (non-) move.

select-next-world (w)

```
var: next-world
if incoherent (w)
then return  $\Phi$  (compute-revisions (w))
else loop for next-world :=
   $\Phi$  (compute-revisions (w)  $\cup$  {w})
  until w = next-world do w := next-world
  finally return next-world
```

6. The Example Revisited

Initially, there would be two preferred diagnoses, $\{fm(\text{computer})\}$ and $\{fm(\text{term}\#1), \dots, fm(\text{term}\#n)\}$, but preference P4

leads to the selection of the initial world

$$w_0 = \{at-most-1-fault\}$$

which avoids the second diagnosis. The only diagnosis under w_0 uniquely determines the computer as defective, hence P1 selects

$$w_1 = \{at-most-1-fault, modg-c_{sub}(computer)\}.$$

Now the internal structure of the computer becomes visible to the diagnosis agent. Because the path from the console terminal to the disk is functional (the operating system runs !), all known fault modes for components along that path are refuted (while those for other components are consistent). Consequently, they are assigned the unknown mode in all preferred diagnoses generated under w_1 . For exactly these components P3 selects the fault-modes-complete hypotheses, fmc .

$$w_2 = \{at-most-1-fault, modg-c_{sub}(computer), \\ fmc(cpu), fmc(mem-ctl), fmc(mem), fmc(bus\#1), \\ fmc(ctl), fmc(disk)\}.$$

The preferred diagnoses are $\{fm(int)\}$ and $\{fm(bus\#2)\}$.

The diagnosis process as described in [Davis 82] proceeds by taking further actions such as structurally refining bus#2, doing a few more probes, switching between alternate models etc. and finally settles on a bent spring-loaded connector.

Provided we had available a diagnosis system that integrated all the necessary diagnosis techniques e.g. alibis, physical negation, physical impossibility, use of non-intermittency and multiple models, generation of test-vectors, probe selection and an appropriate set of preferences, it should now be possible to re-construct the entire process put forward by Davis as a challenge to the AI-community in 1981.

7. Implementation

We have built a system called 'Magellan'¹ that uses the described revision operator. It is applied to the diagnosis of ballast tank systems where the control of multiple, qualitative and quantitative models is crucial [Dressler et. al. 93].

8. Related Work

[Struss 89] provides a framework for *representing* diagnostic hypotheses, a necessary precondition for reasoning with them. Few have followed this avenue since then. An exception is [Nejdl 91] where the notion of 'plausible world' is introduced. A set of such worlds represents a belief state. In contrast to our approach Nejdl's worlds resemble models of the diagnostic situation, i.e. a valid interpretation rather than a set of working hypotheses. Therefore, Nejdl in essence compares different diagnoses while we compare different worlds and thereby sets of diagnoses.

The problem of guiding the diagnostic process is an instance of the more general belief revision problem. This is an active research area within AI, e.g. [Doyle 91], that should have a major impact on diagnosis in the future. Notions, however, like e.g. epistemic entrenchment are important, but not strong enough to provide guidelines for diagnostic processes.

9. Conclusion

We have presented a framework for organizing diagnosis pro-

cesses that exploits and adapts belief revision for diagnostic purposes. Informational economy is the basis for determining candidates for belief state changes of diagnostic agents. Working hypotheses as used in diagnosis systems are the atomic units for building belief bases. Additional diagnostic information allow the agents to make a rational choice not only in terms of an emerging overall belief revision theory but also in terms of preferences specific for diagnosis.

Acknowledgments

Hartmut Freitag, Michael Montag, Peter Struss, Monika Pfau, Wolfgang Nejdl, Thomas Guckenbiehl, Claudia Johnson and Lübke Onken commented on earlier drafts. This work was supported by BMFT, project BEHAVIOR, ITW 9001 A9.

References

- [Böttcher 91] C. Böttcher, Die Verwendung multipler Modelle in der modellbasierten Diagnose, University of Karlsruhe, 1991
- [Dressler et. al. 93] O. Dressler, C. Böttcher, M. Montag, A. Brinkop, Qualitative and Quantitative Models in a Model-based Diagnosis System for Ballast Tank Systems, Proc. ToolDiag'93, Toulouse
- [Console, Torasso 90] L. Console, P. Torasso, Integrating Models of Correct Behavior into Abductive Diagnosis, Proc. ECAI 90, Pitman
- [Davis 82] R. Davis, Expert Systems: Where are we? And Where Do We Go From Here? AI Magazine, Spring 1982
- [de Kleer -Williams] J. de Kleer, B. Williams, Diagnosing Multiple Faults, Artificial Intelligence, 1987
- [de Kleer 91] J. de Kleer, Focusing on Probable Diagnoses, Proc. AAAI 91, MIT Press, 1991
- [Doyle 91] J. Doyle, Rational Belief Revision, Proc. KR '91, Morgan Kaufmann, 1991
- [Dressler 90] O. Dressler, Problem Solving with the NM-ATMS, Proc. ECAI-90, Pitman, 1990
- [Dressler, Farquhar 90] O. Dressler, A. Farquhar, Putting the Problem Solver Back in the Driver's Seat: Contextual Control of the ATMS, in: Springer LNAI 515, 1990
- [Dressler, Struss 92] O. Dressler, P. Struss, Back to Defaults: Characterizing and Computing Diagnoses as Coherent Assumption Sets, Proc. ECAI 92, Wiley & Sons, 1992; Corrected Version in DX-92, Int'l Workshop on Principles of Diagnoses, Seattle, 1992
- [Friedrich et al 90] G. Friedrich, G. Gottlob, W. Nejdl, Physical Impossibility instead of Fault Models, Proc. AAAI 90, MIT Press
- [Gärdenfors 88] P. Gärdenfors, Knowledge in Flux, MIT Press, 1988
- [Hamscher 91] W. Hamscher, Modeling Digital Circuits for Troubleshooting, Artificial Intelligence, Vol 51, 1-3, 1991
- [Nejdl 91] W. Nejdl, Belief Revision, Diagnosis and Repair, Proc. GI-Congress Knowledge-based Systems, Springer 1991
- [Raiman 89] O. Raiman, The Alibi Principle, Proc. Diagnosis Workshop Paris 1989, IBM Report
- [Raiman et al 91] O. Raiman, J. de Kleer, V. Saraswat, M. Shirley, Characterizing Non-intermittent Faults, Proc. AAAI-91, MIT Press
- [Reiter 80] R. Reiter, A Logic for Default Reasoning, Artificial Intelligence, 1980
- [Struss-Dressler 89] P. Struss, O. Dressler, "Physical Negation" - Integrating Fault Models into the General Diagnostic Engine, Proc. IJCAI-89, Morgan Kaufmann
- [Struss 89] P. Struss, Diagnosis as a Process, Working Notes of the International Workshop on Model-based Diagnosis, Paris, 1989, revised version in Readings in Model-based Diagnosis, Morgan Kaufmann, 1992
- [Struss 92] P. Struss, What's in SD? Towards a Theory of Modelling for Diagnosis, Readings in Model-based Diagnosis, Morgan Kaufmann, 1992

1. Fernão de Magalhães (1480-1521), portuguese sailor