

Diagnosing Temporal Misbehavior

Gerhard Friedrich and Franz Lackinger

Christian Doppler Laboratory for Expert Systems

Vienna University of Technology, Austria

(friedrich or lackinger)@vexpert.dbai.tuwien.ac.at

Abstract

One of the main goals of model-based diagnosis is the detection and localization of failures that are very difficult to diagnose even for human experts. Especially transient failures require a sophisticated problem-solving strategy and often overcharge an experienced human diagnostician. Unfortunately, such transient failures occur in many applications and cause unacceptable misbehavior. Therefore, dealing with failures which have a temporal aspect is a challenging task for model-based diagnosis.

However, current research on model-based diagnosis treats assumptions about the component behavior as static and therefore fails to reason about time-varying failures. In this paper we present a new approach to model-based diagnosis which enables us to deal with its temporal aspects.

1 Introduction

Reasoning about behavioral abnormality of technical devices is a key problem for model-based diagnosis. A lot of research work has already been done in this field which has typically concentrated on static devices and permanent abnormality (e.g. [Friedrich *et al.*, 1990b; Davis, 1984]). Only recent approaches deal with the temporal aspects of abnormality (e.g. [Hamscher, 1991; Raiman *et al.*, 1990]).

Unfortunately, most practical applications require the ability of a diagnosis engine to detect and localize transient failures. Since the early days of fault-tolerant computing most researchers in this field agree that diagnosis of transient failures is of crucial importance.

For example, [Breuer, 1973] writes

"Given that intermittents are tremendously more important to diagnose and that they do occur frequently, I propose that finding an orderly means to diagnose them is the most important unsolved problem in this area today."

In this paper we propose a framework for reasoning about the temporal aspect of misbehavior and therefore extend the traditional approach to model-based diagnosis. As a result, we are able to diagnose all types of

failures including environmental disturbances as well as physical damages.

In Section 2 we start with a discussion of various problems of reasoning about temporal abnormality. Additionally, we give a clear distinction between failures and faults and classify them according to various properties. In Section 3 we present our framework of model-based diagnosis introducing *failure intervals* and *temporal diagnoses*. We introduce a probabilistic approach to assign plausibility values to diagnoses which is based on the theory of Markov-processes in Section 4. Finally, in Section 5 we compare our framework to related research work in this field and present topics for future work.

2 Classifying Temporal Misbehavior

We start the discussion with the presentation of an example which illustrates the problems of reasoning about various classes of faulty behavior and their causes. This discussion will lead us to the requirements which a model-based diagnosis engine has to fulfill if it is applied to time-varying systems.

Consider a proportional-control heater/cooler [Ng, 1990]. This device consists of five components - the temperature sensor s , the temperature control knob k , the power supply o , the electric heating/cooling component e and the proportional controller c . With respect to the difference (denoted with d) between the observed actual room temperature (t_i^{ob}) and the observed intended room temperature t_s^{ob} (which is set by k), e generates heat flow hf_i which, depending on the direction of the difference, can either be positive or negative. The adjustment amount is denoted with ehf_i , the (observed) power input to the heater/cooler is denoted with $(p^{ob}) p$.

Figure 1 shows the constraint model which associates variables to components (we follow [Kuipers, 1986] in our terminology):

The qualitative quantity space for each variable is depicted in Table 1 ($minf < cold < t_{mid} < t_{room} < inf$, $minf < 0 < inf$, $0 < p_{max} < inf$).

In the following discussion we will use our example to discuss various scenarios of faulty behavior. We denote the qualitative system states as sets of qualitative parameter values and create a scenario as a sequence of system states.

We distinguish between faults from two viewpoints

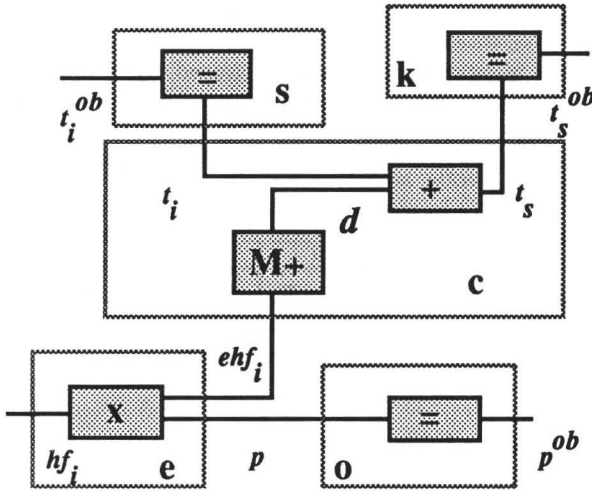


Figure 1: Constraint Network (Heater/Cooler)

$$\begin{aligned}
t_i^{ob} &= cold \vee t_i^{ob} \in (cold, t_{mid}) \vee t_i^{ob} = t_{mid} \vee t_i^{ob} \in (t_{mid}, t_{room}) \\
\vee t_i^{ob} &= t_{room} \vee t_i^{ob} \in (t_{room}, inf) \vee t_i^{ob} = inf. \\
t_i &= cold \vee t_i \in (cold, t_{mid}) \vee t_i = t_{mid} \vee t_i \in (t_{mid}, t_{room}) \\
\vee t_i &= t_{room} \vee t_i \in (t_{room}, inf) \vee t_i = inf. \\
t_s &= t_{room} \vee t_s \in (t_{room}, inf) \vee t_s = inf. \\
t_s^{ob} &= t_{room} \vee t_s^{ob} \in (t_{room}, inf) \vee t_s^{ob} = inf. \\
hf_i &= minf \vee hf_i \in (minf, 0) \vee hf_i = 0 \\
\vee hf_i &\in (0, inf) \vee hf_i = inf. \\
ehf_i &= minf \vee ehf_i \in (minf, 0) \vee ehf_i = 0 \\
\vee ehf_i &\in (0, inf) \vee ehf_i = inf. \\
p^{ob} &= 0 \vee p^{ob} \in (0, p_{max}) \vee p^{ob} = p_{max} \\
\vee p^{ob} &\in (p_{max}, inf) \vee p^{ob} = inf. \\
p &= 0 \vee p \in (0, p_{max}) \vee p = p_{max} \vee p \in (p_{max}, inf) \vee p = inf.
\end{aligned}$$

Table 1: Quantity Spaces

using a *temporal classification* and a *value classification*. Our terminology is closely related to the terms which have been used by the *Fault-Tolerance* community as a standard for the last decade [Kopetz, 1982; Avizienis, 1982].

According to the distinction between *failure* and *fault* used by that community, we propose the notion *fault* to denote that the system's behavior deviates from the specification of correct behavior for some time interval. Note that the terms *symptom* and *misbehavior*, also used by the AI-Diagnosis community, are equivalent to our notion *fault*. The notion *failure* denotes the causes of faults. This distinction is important with respect to the temporal classification introduced in the following.

2.1 Temporal Classification

Figure 2 depicts the relation between failures and their possible manifestations (i.e. a failure can cause permanent or transient faults). Nevertheless, failures need not necessarily alter the correct behavior during the mission of the system.

Failures have a temporal aspect and they do not necessarily cause faulty behavior (e.g. their extent might be held locally by means of fault-tolerance). Conversely, every fault is caused by at least one failure. In order to distinguish between internal and external causes of faults, we introduce the notions *defects* for the former

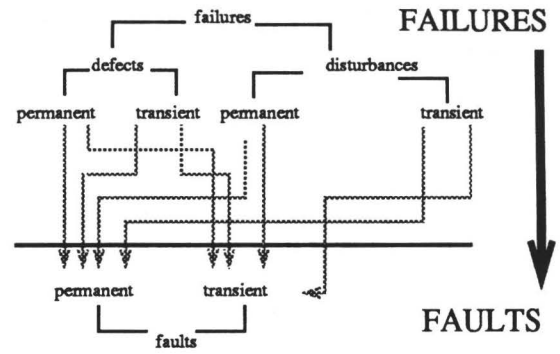


Figure 2: The Relation Between Failures and Faults

and *disturbances* for the latter. This distinction enables us to decide whether the system has to be repaired or the environmental conditions must be changed in the course of re-establishing correct behavior. Not also that the concept *defect* covers physical damages as well as design and assembling failures.

2.1.1 Permanent Faults

Let us assume that the proportional controller has a physical defect (e.g. component *c* exhibits a *stuck-at-0* fault at *d*). This defect exhibits the behavior pattern depicted in Table 2.

	s0	s1	s2	s3
t_i	t_{room}	$(cold, t_{room})$	$cold$	$cold$
t_s	t_{room}	t_{room}	t_{room}	t_{room}
d	0	0	0	0

Table 2: Permanent Faults

It is consistent to assume in state s0 that the system behaves correctly. However, from state s1 to state s3, it is continuously inconsistent to assume that the system behaves correctly. The crucial property of a permanent fault is its permanent duration, not its beginning (i.e. a permanent fault can start *during* diagnosis).

Note that although the abnormal behavior is permanent, its cause can be temporally limited (i.e. transient). For example, a temporal disturbance which changes the contents of a counter can cause a clock to behave continuously faulty.

2.1.2 Transient Faults

In order to introduce transient faults we discuss two examples in more detail. In addition to physical defects we present an example for an environmental disturbance because this type of failure is a prominent cause for transient faults.

Physical Defects: Let us now assume that a physical defect in the sensor causes a temporal deviation of the room temperature recorded by the sensor (t_i) from the actual room temperature (t_i^{ob}) (e.g. the response of the sensor is too slow due to an internal time delay) (see Table 3). Although the system shows transiently faulty behavior, its cause can permanently reside.

	s0	s1	s2	s3	s4	s5
t_i^{ob}	t_{room}	$(cold, t_{room})$	$cold$	$cold$	$cold$	$cold$
t_i	t_{room}	(t_{mid}, t_{room})	t_{mid}	$(cold, t_{mid})$	$cold$	$cold$

Table 3: Transient Faults - Example 1

In contrast to permanent abnormality, this temporal deviation from the specified correct behavior is only visible for a finite period of time (during states s2 - s3). In state s4, however, the fault has disappeared, although the physical defect might still exist. Consequently, we can assume that the system behaves correctly in the states s4 and s5. Nevertheless, if we start with the same initial conditions we will get the same result (i.e. the fault is reproducible).

Environmental Disturbances: In the previous example we have considered abnormality which can causally be explained by physical defects in the system.

However, each physical system works in an environment which influences its behavior in an unpredictable manner. For example, alpha particles can alter the state of memory cells. Temperature and humidity influence the correct behavior of electronic circuits.

From the viewpoint of repair, it is very important to distinguish between physical defects and environmental disturbances. The former have to be repaired, while the latter can only be tolerated in most cases. Therefore we have to include this distinction into our diagnosis process.

For example, consider the behavior pattern in Table 4.

	s0	s1	s2	s3	s4
ehf_i	$(0, inf)$	$(0, inf)$	$(0, inf)$	$(0, inf)$	$(0, inf)$
hfi	$(0, inf)$	$(0, inf)$	0	$(0, inf)$	$(0, inf)$
p	p_{max}	$(0, p_{max})$	0	$(0, p_{max})$	p_{max}
p^{ob}	p_{max}	p_{max}	p_{max}	p_{max}	p_{max}

Table 4: Transient Faults - Example 2

This transient fault is caused by a temporary external disturbance (e.g. the environmental temperature is too high) that results in a temporary shutdown of the power supply. Note that this fault is not caused by a component-oriented failure – only the observed behavior deviates from the desired. In fact, we have to concentrate on the causes of abnormality (e.g. physical defects or environmental disturbances) in order to repair.

2.2 Value Classification

In the above discussion we have assumed determinate faulty behavior. Determinate faults (also called *stuck-at* faults) do not show arbitrary behavior, the faulty behavior is reproducible. More precisely, determinate behavior is characterized by a functional relation between the exogenous and the endogenous variables of the system.

Unfortunately, some failures do not cause determinate behavior. In this sense, the impact of an alpha particle on a counter produces indeterminate, transiently faulty behavior. For example, let us assume the following two behavior patterns of the heater/cooler that are depicted

in Table 5 (state sN denotes a possible successor state of s0).

This indeterminate fault is caused by a loose cable in the sensor display which produces different values of resistance. Note that these behavior patterns are only two examples from a set of possible faults resulting from this failure.

	s0	sN
t_i^{ob}	t_{room}	t_{room}
t_i	t_{room}	$cold$
t_i^{ob}	t_{room}	t_{room}
t_i	t_{room}	t_{mid}

Table 5: Indeterminate Fault

2.3 Requirements for Temporal Diagnosis

From the above discussion we derive the following conclusions and requirements for reasoning about temporal abnormality. These requirements serve as a guideline for our approach to model-based temporal diagnosis.

In order to distinguish between temporal properties of failures we have to model them explicitly. This is due to the fact that we can not draw conclusions about the temporal nature of failures by *solely considering the temporal properties* of their manifestations (i.e. the faults). For example, transient faults can be caused by permanent and transient failures both.

It is not sufficient to distinguish between determinate and indeterminate behavior of components for making assumptions about permanent and transient behavior (i.e. intermittent, according to the terminology of [Raiman *et al.*, 1990]). We rather have to reason about the possible transitions between different combinations of failures. This view enables us to distinguish between the different classes of failures with a finer temporal granularity and leads to an enhanced characterization of the assumptions about the system's history and its present state (i.e. diagnoses).

The operation of every device depends significantly on its environmental conditions. Consequently, we have to include the environment in our models to be able to reason about it. This is a necessary presupposition to distinguish between various causes for faulty behavior (i.e. defects and disturbances).

3 Defining Failures and Temporal Diagnoses

Continuing the above discussion, we present our new approach to model-based temporal diagnosis.

We start with a formal definition of *failures* and their temporal properties. Furthermore, we extend the traditional notion of a diagnosis with respect to these properties and introduce *temporal diagnoses*.

3.1 Failures

Faults represent the abnormal behavior of components and systems, whereas failures are the causes of such abnormal behaviors.

Definition 1 (Failures) FAI is a finite set of constants representing failures.

The set of possible failures for a component c is denoted by $FAI(c)$. A failure is associated with a component c at time instance t by the predicate *failure*, e.g. $failure(c, shorted)_t$. This association allows a classification which is valid for a broad range of failures. Nevertheless, we do not treat this association as a necessary presupposition but rather as an aid.

Consequently, we distinguish between component-dependent failures, which we denote *defects*, and component-independent failures which we call *disturbances*.

Note that a failure may not be in contradiction with the correct behavior and that assuming a component to be associated with several failures at time point t is valid. For example, $\neg ab(c) \wedge failure(c, output_shorted)_t$ as well as $failure(c, output_shorted)_t \wedge failure(c, input_shorted)_t$ are no contradictions ($\neg ab(c)$ denotes the correct *behavior* of component c).

This discussion leads us to the notion *failure behavior*.¹

Definition 2 (Failure Behavior) The behavior of a failure $f \in FAI$ for an entity $c \in COMPS$ is a set of First-Order-Sentences B , such that $\forall T : failure(c, f)_T \rightarrow B$.

According to this definition, a fault can be characterized by several failure behaviors. [Hamscher, 1991] introduces the concept of a *syndrome* which denotes sets of failures having the same behavior. This concept reduces the complexity of the diagnosis process since the number of failures can be significantly diminished. This idea can easily be integrated into our framework by viewing a syndrome as an abstract failure.

In contrast to related approaches which are based on behavioral component modes, we view the central goal of model-based diagnosis to identify failures. Consequently our concept of a *temporal diagnosis*, as defined later, will be a set of assumptions about the existence and absence of failures.

Similar to the unknown fault mode of [de Kleer and Williams, 1989; Struss and Dressler, 1989] we include unknown failures if the failure description is not complete.

3.1.1 Temporal Classification

We distinguish between permanent and transient failures to make a first qualitative distinction between different levels of plausibility.

In order to define temporal diagnoses we have to express temporal properties of failures. We therefore define failure intervals as follows. Let $in(T, Sets)$ be an abbreviation for $\bigvee_{s \in Sets} T \in s$.

Definition 3 (Failure Intervals) $INTFAI(c, f)$ is the finite set of closed time intervals for a entity $c \in COMPS$ and a failure $f \in FAI$ such that

- For each pair of different intervals $[A_1, E_1], [A_2, E_2]$ the relation holds that one interval is before the other, i.e.

$$\begin{aligned} \forall A_1, E_1, A_2, E_2 : & [A_1, E_1] \in INTFAI(c, f) \wedge \\ & [A_2, E_2] \in INTFAI(c, f) \wedge \\ & [A_1, E_1] \neq [A_2, E_2] \\ & \rightarrow \exists t : (E_1 < t < A_2) \vee \\ & (E_2 < t < A_1). \end{aligned}$$

- Entity c is assumed to have failure f :

$$\forall T : in(T, INTFAI(c, f)) \leftrightarrow failure(c, f)_T.$$

The sentence $\forall T : in(T, INTFAI(c, f)) \leftrightarrow failure(c, f)_T$ expresses that at all time points which are included in an interval $i \in INTFAI(c, f)$ the component c has failure f . At all time points which are in no failure interval for f the component c does not have failure f . If we want to express that a component c does not have failure f then $INTFAI(c, f) = \{\emptyset\}$.

For example, $INTFAI(r_1, shorted) = \{[t1, t2], [t3, t4]\}$ has the meaning that entity r_1 has the failure *shorted* exactly in the time intervals $[t1, t2]$ and $[t3, t4]$.

According to the above discussion, we formalize the notions of *permanent* and *transient failure*.

Permanent Failures: Permanent failures are irreversible changes in components, such as shorted transistors or continuous environmental disturbances. They lead to a permanent transformation of the original system into a new system that has a different description and will not always behave in the proper manner.

Let t_Ω be the time point when the system has fulfilled its purpose and t_α be the opening of operations of the system (i.e. the birth of the system).

Definition 4 (Permanent Failure) A failure $f \in FAI$ for a component $c \in COMPS$ and a relevant interval $[t_\alpha, t_\Omega]$ is permanent starting with t_b iff there exists a time interval $[t_b, t_e] \in INTFAI(c, f)$ such that $t_b \leq t_\Omega \leq t_e$.

Transient Failures: The attribute *transient* is assigned to a failure which resides only temporarily in the system, but which can nevertheless cause permanent faults. An obvious cause of such faults is external interference of short duration, for example, the impact of an alpha particle on a memory cell. Another cause for a temporary deviation of component variables are temperature variation or delays in timing caused by stray capacitance.

Definition 5 (Transient Failure) A failure $f \in FAI$ for a component $c \in COMPS$ and the relevant interval $[t_\alpha, t_\Omega]$ is transient iff there exists a time interval $[t_b, t_e] \in INTFAI(c, f)$ such that $t_e < t_\Omega$.

Note that these definitions cover a broad range of possible combinations of failures. For example, we are able to represent the frequent case that transient failures can start to be permanent later.

3.2 Temporal Diagnosis

According to the above concepts we define a temporal diagnosis as follows. Let $FS = \{(c, f) | c \in COMPS, f \in FAI(c)\}$.

¹In the following $SD, COMPS$, and OBS are defined as usually (e.g. [Reiter, 1987]). However, $COMPS$ is extended in order to include constants representing the environment, the design, etc., in addition to device components. Therefore we rather talk about entities than about components. Additionally, SD and OBS may be time-varying.

Definition 6 (Temporal Diagnosis) A temporal diagnosis Δ for a relevant interval $[t_\alpha, t_\Omega]$ is a set of triples $\{(c, f, INTFAI(c, f)) \mid (c, f) \in FS\}$ for $(SD, COMPS, OBS, FAI)$ such that

$$SD \cup OBS \cup \{\forall T \in [t_\alpha, t_\Omega] : in(T, INTFAI(c, f)) \leftrightarrow failure(c, f)_T \mid (c, f, INTFAI(c, f)) \in \Delta\}.$$

is satisfiable.

A temporal diagnosis assigns to each failure (which is associated to an entity) a set of failure intervals, such that, with respect to the relevant interval, in these failure intervals the failure is present and at all time points not covered by these intervals the failure is *not* present.

This definition of a temporal diagnosis includes all consistent assumptions about possible occurrences of failure combinations. Furthermore, the concept of a temporal diagnosis is general enough to allow the application of general theorem provers for consistency check (e.g. [Friedrich and Nejd, 1990]).

We can apply various criteria how to assign plausibility values to the diagnoses. In Section 4 we present techniques for the computation of such plausibility values based on the failure rate λ and the repair rate μ .

Note that we do not need a temporal classification of *behavior* (i.e. faults) for temporal diagnosis because we can derive faults from failures. For a given temporal diagnosis we are able to predict this faulty behavior weighted with a plausibility value.

As a result, we have achieved a straightforward generalization of the traditional view of diagnosis by adding a temporal dimension to the candidate space (i.e. diagnosis lattice).

4 Ranking of Temporal Diagnoses

We can distinguish between various criteria for the plausibility of a diagnosis. The most popular ones are either the minimal or irreducible or most probable sets of assumptions about the behavior of components.

The diagnosis process as introduced by [de Kleer and Williams, 1989] generates the set of most probable (leading) diagnoses. The different modes of behavior of the components are weighted by a static *a priori* probability value. It is assumed that components are in a specific mode of behavior and do not switch between various modes during diagnosis. The *a posteriori* probabilities are computed for these diagnoses in order to rank them and to select the best next measurement point.

If, however, we include time into diagnosis we have to allow transitions between states when failures exist and when they have vanished.

We therefore propose the use of an appropriate statistical model using Markov-processes to compute probabilities of temporal diagnoses.

4.1 Statistical Model

We presuppose that failures are independent. Furthermore, time is assumed to be linear and discrete and will be represented by natural numbers.

Let $P_f(t)$ denote a probabilistic function ($P_f(t_0) = 1$, $P_f(\infty) = 0$) where $P_f(t)$ is the probability that failure f is not present in the system until time point t .

Let further $d_f(t)$ denote the failure density function for failure f with

$$d_f(t) = -\frac{dP_f(t)}{dt}$$

We define the failure rate $\lambda_f(t)$ for failure f as

$$\lambda_f(t) = \frac{d_f(t)}{P_f(t)}$$

Based on a probabilistic function $M_f(t)$ which expresses the probability that failure f has been repaired until time point t , the repair rate $\mu_f(t)$ is defined analogously to $\lambda_f(t)$. $\mu_f(t)$ covers self-repair and external repair.

For many failures $P_f(t)$ and $M_f(t)$ will be exponential functions and therefore result in a constant failure rate and repair rate. This is valid for most systems during the main part of their life cycle, excluding short time intervals after the start and before the end of the operation. Therefore, we assume both λ and μ to be constant in the further discussion.

In the following we show how to compute the statistics of temporal diagnoses using Markov processes. In our approach, each state of a Markov process represents a set of assumptions about the presence or absence of each failure $f \in FAI$. The transition arc between two nodes represents one discrete step from time point t_i to time point t_{i+1} .

4.2 Computation

Let Δ be an element from the set of temporal diagnoses and let $f \in FAI$ be a failure. Let t_0 be a time point where no failure is present in the system (i.e. $t_0 = t_\alpha$). Consequently, the initial probability at time point t_0 that f is not present is 1. In order to compute the probabilities for a temporal diagnosis at time point t_n (i.e. $t_n = t_\Omega$), the following computation step has to be applied to all time points t_i for $i = 1, \dots, n$.

The temporal diagnosis Δ defines for each time point t_i a partition of FAI into the two disjunct subsets FAI_i^p and FAI_i^n . FAI_i^p (FAI_i^n) denotes the set of failures which are assumed to be present (absent) at time point t_i .

We distinguish between four types of transitions between time point t_i and t_{i+1} . $p_{t_i}(\Delta)$ denotes the actual probability of Δ at time point t_i .

We compute $p_{t_{i+1}}$ as follows:

Let $p_{t_{i+1}}(\Delta) := p_{t_i}(\Delta)$.

1. **fail transition:** for each failure $f \in FAI_i^n$ and $f \in FAI_{i+1}^p$:
 $p_{t_{i+1}}(\Delta) := p_{t_i}(\Delta) \times \lambda_f$
2. **no-fail transition:** for each failure $f \in FAI_i^n$ and $f \in FAI_{i+1}^n$:
 $p_{t_{i+1}}(\Delta) := p_{t_i}(\Delta) \times (1 - \lambda_f)$
3. **repair transition:** for each failure $f \in FAI_i^p$ and $f \in FAI_{i+1}^n$:
 $p_{t_{i+1}}(\Delta) := p_{t_i}(\Delta) \times \mu_f$

4. **no-repair transition:** for each failure $f \in FAI_t^p$ and $f \in FAI_{t+1}^p$:

$$p_{t+1}(\Delta) := p_{t+1}(\Delta) \times (1 - \mu_f)$$

In order to normalize the probabilities of all temporal diagnoses with respect to the interval $[t_0, t_i]$, we divide their probabilities by the sum of the probabilities of all temporal diagnoses in $[t_0, t_i]$.

However, for realistic applications the large number of possible diagnoses will prevent the computation of the normalization. As noted by [de Kleer and Williams, 1990], this normalization does not affect the ranking of the diagnoses.

In model-based diagnosis, we distinguish between diagnoses which explain the observations, and diagnoses which do not constrain the possible outcome of a probe. From a rational point of view, such *observation-explaining* diagnoses are more plausible and therefore preferred. In [de Kleer and Williams, 1990], the probabilities of these diagnoses are unchanged, while the probabilities of not explaining diagnoses are divided by the number of possible outcomes of the probe. In particular, the concept of *explanation* is related to the outcome of the next probe. The value that expresses the capability of a diagnosis to explain outputs (i.e. endogenous) is only recomputed if additional inputs (i.e. exogenous) are selected and measured.

Similar to [de Kleer and Williams, 1990], we use Bayes' rule to compute the posterior probability of a temporal diagnosis and also assume that the set of possible outcomes of a probe is finite. We extend the heuristic of [de Kleer and Williams, 1990] to weight the explanatory capabilities of a temporal diagnosis:

Each temporal diagnosis divides OBS into the two disjunct subsets OBS_{en} (denoting endogenous observations) and OBS_{ex} (denoting exogenous observations). We only have to consider the explanatory capability of temporal diagnoses with respect to OBS_{en} , because exogenous observations can only be explained by the environment.

For each $obs \in OBS_{en}$ of a temporal diagnosis Δ we weight the probability $p_{t_n}(\Delta)$ as follows:

$$p_{t_n}(\Delta) := p_{t_n}(\Delta) \times w \begin{cases} w = 1 & \text{if } SD \cup \Delta \cup (OBS \setminus obs) \vdash obs \\ w = \frac{1}{m_{obs}} & \text{otherwise} \end{cases}$$

where m_{obs} is the number of possible outcomes of a probe which has been observed to be obs .

4.3 Example

In this section we show how our approach applies to the example of Section 2.

We consider the subsystem that consists of the power supply o and the heating/cooling component e . Consequently, we observe the two endogenous variables p and hf_i (the exogenous variables are $(p^{ob} = p_{max})_T$ and $(ehf_i = (0, inf))_T$ for all $T \geq t_1$).

As an example, two possible failures are included:

$$\forall T: failure(o, f_o)_T \rightarrow (p = 0)_T.$$

$$\forall T: failure(e, f_e)_T \rightarrow (hf_i = 0)_T.$$

with $\lambda_{f_o} = 1 \times 10^{-2}$, $\lambda_{f_e} = 1 \times 10^{-2}$, $\mu_{f_o} = 1 \times 10^{-5}$, $\mu_{f_e} = 1 \times 10^{-5}$.

The history of the system, containing the endogenous observations and the weighted transitions, is depicted in Figure 3 (f_o (f_e) means this failure is in the system). As stated above, we assume that at time point t_0 no failure is present in the system.

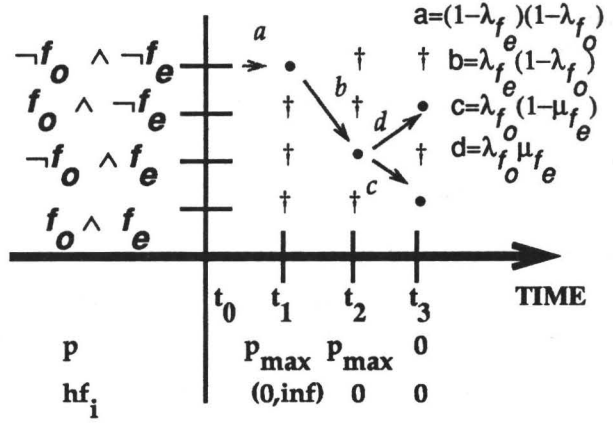


Figure 3: Possible Failure Combinations and Their Transitions

In Figure 3, bullets denote the nodes of the Markov process while daggers denote combinations of failure assumptions that are not in any temporal diagnosis.

Table 6 depicts the weighted temporal diagnoses (not normalized, both diagnoses are explaining).

TRANS.	DIAGNOSES	PROB.*
$\langle a, b, c \rangle$	$\langle (o, f_o, [t_3, t_3]), (e, f_e, [t_2, t_3]) \rangle$	9.6×10^{-5}
$\langle a, b, d \rangle$	$\langle (o, f_o, [t_3, t_3]), (e, f_e, [t_2, t_2]) \rangle$	9.7×10^{-10}

*probability not normalized

Table 6: Temporal Diagnoses

The first diagnosis in Table 6 is more likely since the probability that failure f_e is persistent is higher than the probability of its repair.

5 Related Work

Beside others, there are at least two remarkable papers (i.e. [Genesereth, 1984] and [Raiman et al., 1990]) dealing with transient (intermittent) behavior. In the terminology of [Raiman et al., 1990], a component behaves *non-intermittently* if its outputs are a function of its inputs, e.g. the additional axiom for an or-gate is:

$$\forall T \text{ ni}(or) \equiv [out(or, T) = F(in_1(or, T), in_2(or, T))].$$

where F represents an unspecified function.

Since the temporal aspect of the faulty behavior of components is not represented, the only achievable statement about defect components is whether they behave *intermittently* or not, no matter whether the current behavior status is *correct* or *abnormal*. Clearly, one of our goals is to express exactly *when* the system behaves correctly and abnormally.

In order to use non-intermittency axioms the non-intermittent behavior (correct or faulty) has to be functional. Consequently, the general I/O-behavior has to be modeled functionally even for permanent failures which is a severe restriction of generality. If we introduce such non-intermittency axioms, random number generators have intermittent behavior, for example.

There are some other important papers extending static diagnosis (i.e. [Hamscher, 1991]). Their work focuses on dealing efficiently with temporal system descriptions and considers the assumptions about failures as permanent whereas our aim is to deal with the temporal properties of diagnoses.

6 Conclusions and Future Work

We have presented a new framework for model-based diagnosis which overcomes many of the current limitations. The distinction between failures and faults and their classification according to *time* and *value* has proven to be important in the course of diagnosing time-varying systems.

We have introduced *temporal diagnoses* as a generalization of the traditional notion of a diagnosis by a temporal aspect. This generalization enables us to detect and localize permanent and transient failures.

Furthermore, we have presented a probabilistic approach using Markov-processes to assign plausibility values to temporal diagnoses. The use of Markov-processes allows us to include failure - and repair - rates and to evaluate all types of transitions between possible failure states.

We are currently investigating various search techniques (heuristic) for plausible temporal diagnoses using physical impossibility [Friedrich *et al.*, 1990a] as one of several pruning techniques. The problem of finding the most plausible diagnoses is one of the most challenging and open problems in current model-based diagnosis research.

Acknowledgements

The contents of this paper have benefited from discussions with Anton Beschta, Oskar Dressler, Hartmut Freitag, Michael Montag, and Peter Struss.

References

- [Avizienis, 1982] Algirdas Avizienis. The four-universe information system model for the study of fault-tolerance. In *12th International Symposium on Fault-Tolerant Computing FTCS-12*, pages 6–13. published by IEEE, June 1982.
- [Breuer, 1973] Melvin A. Breuer. Testing for intermittent faults in digital circuits. *IEEE Transactions on Computers*, C-22(3):241–246, March 1973.
- [Davis, 1984] Randall Davis. Diagnostic reasoning based on structure and behaviour. *Artificial Intelligence*, 24:347–410, 1984.
- [de Kleer and Williams, 1989] Johan de Kleer and Brian C. Williams. Diagnosis with behavioral modes. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1324–1330, Detroit, August 1989. Morgan Kaufmann Publishers, Inc.
- [de Kleer and Williams, 1990] Johan de Kleer and Brian C. Williams. Focusing the diagnosis engine. In *First International Workshop on Principles of Diagnosis*, Stanford, July 1990.
- [Friedrich and Nejd, 1990] Gerhard Friedrich and Wolfgang Nejd. MOMO — Model-based diagnosis for everybody. In *Proceedings of the IEEE Conference on Artificial Intelligence Applications (CAIA)*, Santa Barbara, March 1990.
- [Friedrich *et al.*, 1990a] Gerhard Friedrich, Georg Gottlob, and Wolfgang Nejd. Physical impossibility instead of fault models. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 331–336, Boston, August 1990.
- [Friedrich *et al.*, 1990b] Gerhard Friedrich, Franz Lackinger, and Wolfgang Nejd. Redefining the candidate space in model-based diagnosis. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 277–282, Stockholm, August 1990.
- [Genesereth, 1984] M. R. Genesereth. The use of design descriptions in automated diagnosis. *Artificial Intelligence*, 24:411–436, 1984.
- [Hamscher, 1991] Walter Hamscher. Modeling digital circuits for troubleshooting. *Artificial Intelligence*, 1991. to appear.
- [Kopetz, 1982] Hermann Kopetz. The failure fault (ff) model. In *12th International Symposium on Fault-Tolerant Computing FTCS-12*, pages 14–17, June 1982.
- [Kuipers, 1986] Benjamin Kuipers. Qualitative simulation. *Artificial Intelligence*, 29:289–388, 1986.
- [Ng, 1990] Hwee Tou Ng. Model-based, multiple fault diagnosis of time-varying, continuous physical systems. In *Proceedings of the IEEE Conference on Artificial Intelligence Applications (CAIA)*, pages 9–15. IEEE Computer Society Press, 1990.
- [Raiman *et al.*, 1990] Olivier Raiman, Johan de Kleer, and Vijay Saraswat. Exploiting non intermittency. In *First International Workshop on Principles of Diagnosis*, pages 120–123, Stanford, July 1990.
- [Reiter, 1987] Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.
- [Struss and Dressler, 1989] Peter Struss and Oskar Dressler. Physical negation — Integrating fault models into the general diagnostic engine. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1318–1323, Detroit, August 1989. Morgan Kaufmann Publishers, Inc.