What's in SD?
Towards a Theory of Modeling for Diagnosis

Peter Struss
Technical University of Munich, Germany
struss@informatik.tu-muenchen.de

Abstract
This paper presents elements of a theory of multiple modeling for consistency-based diagnosis that provide a foundation for addressing the essential problems:
- Characterize relations between different models in order to determine their role in diagnosis and the impact a model switch has on the space of diagnoses.
- Determine criteria for deciding when to switch to which model.
Using a "real-life" example from the domain of power transmission networks, we define and illustrate some basic logical relations among models and describe their impact on the space of diagnoses. The results lay the foundation for a control strategy that starts off with the simplest and cheapest models, and that retracts modeling assumptions and switches to more detailed models guided by information obtained from the diagnostic process itself.

1 Introduction and Outline

Diagnosis of technical systems involves determining the actual modes the constituent parts of the system are working in (the correct mode, or one of the known faults, or some unknown behavior). Model-based diagnosis does so by using models of the behavior modes. Consistency-based diagnosis uses models for determining behavior modes by ruling out (combinations of) behavior modes, if the respective models are contradicting the actual system's behavior. Obviously, a theory of model-based diagnosis has to be a theory of the content and diagnostic use of models in the first place.

However, as a more or less explicit research strategy, most of the approaches focused on problems of the diagnostic procedure, presuming there exists a simple and unique way of modeling the device or avoiding to specify its nature. More structure has been imposed on the model by incorporating fault models ([de Kleer-Williams 89], [Hamscher 90], [Holtzblatt 88], [Struss 88b], [Struss-Dreasler 89]) and hierarchy ([Davis 84], [Hamscher 90], [Struss 88a, 89b]). Like a hierarchical structure, the distinction between different views on component models ([Davis 84], [Struss 88a,b], [Hamscher 90]) aims at exploiting different perspectives and at reducing the complexity of diagnosis of non-trivial artifacts.

Although considered as an essential ingredient to model-based reasoning from the very beginning ([Davis 82]), some of the most powerful means for this purpose, namely reasoning with abstractions, simplifications and approximations, has not been deeply investigated and exploited so far, and moreover, we lack a coherent theory which addresses all these issues and which fits the formalisms of model-based diagnosis.

Why is this difficult? Generally speaking, abstractions may turn out to be too coarse to achieve a satisfactory diagnosis, and simplifications may miss important features of a certain problem. Trivially, if a diagnostic system uses a model that is inadequate for a particular case at hand, the resulting diagnosis is likely to be wrong or at least useless; the model may fail to reveal an existing contradiction, or it may suggest an inconsistency that is only virtual. The dilemma is: the best protection against such failures, namely using the best, most accurate and most detailed model available, tends to make the task intractable, at least unnecessarily complex for the "simpler cases".

And what makes the system work efficiently, namely certain assumptions that simplify the model, may render it useless if these assumptions are violated in the situation to be dealt with. This is one reason why we would like to have multiple models of a system and let the diagnostic system use the one appropriate for the particular case. It should also

* Extended version of a presentation at the 2nd International Workshop on Principles of Diagnosis, Milan, Italy, 1991
be appropriate w.r.t. the stage of the diagnostic process.

The guiding principle for selecting models ought to be "Use models as simple as possible and as sophisticated as necessary". It implies the model might be exchanged during the diagnostic process. First, because it may not be obvious from the very beginning what model is required for the particular case. Second, because different stages in the process require different models (e.g. abstract and cheap models for generating initial diagnostic candidates and more detailed ones for discriminating among them). Switching to a different model should not mean merely starting the procedure again from zero, but allow for carry over of as much information as possible from previous steps. This requirement sets the essential problems we have to solve:

- Characterize relations between different models in order to determine their role in diagnosis and the impact a model switch has on the space of diagnoses.
- Determine criteria for deciding when to switch to which model.

This paper presents elements of a theory of multiple modeling for consistency-based diagnosis that provide a foundation for a solution to the aforementioned problems. One focus is on the crucial distinction between what we will call a view and a simplification of models:

- In the first case, one model strictly implies the other one, whilst
- a simplification of a model is a proper substitute only if certain simplifying assumptions are satisfied.

We treat such modeling assumptions as instances of "working hypotheses" and, in section 3, briefly introduce DP (Diagnosis as a Process, [Struss 89a]), an extension of consistency-based diagnosis that enables diagnosis under retractable diagnostic assumptions.

A rigorous formal treatment of relations between models requires a rigorous, formal theory of models themselves and of the general concept of a model. Many model-based diagnosis systems describe behavior of physical systems in terms of constraints on values of local variables and parameters, i.e. by (mathematical) relations. In [Struss 91] we presented a formal theory of relational models that reflects the requirements of consistency-based diagnosis and allows us to precisely distinguish between different types of model transformations, such as abstraction and simplification. This theory is summarized and illustrated in section 4.

Section 5 defines and illustrates some basic logical relations among models and describes their impact on the space of diagnoses. The results lay the foundation for control strategies that start off with the simplest and cheapest models, and that retract modeling assumptions and switches to more detailed models guided by information obtained from the diagnostic process itself.

Such a strategy is introduced in section 6 and illustrated by an example taken from the application domain of DPNet, a diagnosis system for power transmission networks ([Beschta et al. 92]), which has been implemented as an instance of GDE7 ([Struss-Dressler 89]). As we will use this domain as a source of examples throughout the paper, an introduction to this application is given in the following section.

Readers who are interested in the general theory only (and who think they will associate the right intuition without illustrations) may skip this section as well as the examples presented in sections 4 through 6.

2 Fault Localization in Power Transmission Networks

2.1 The Problem

What is the problem to be tackled? The purpose of a power transmission network is connecting a number of sinks (potentially transformers to a lower voltage level) to operating sources (also possibly transformers). This has to be done in a way that guarantees energy transmission also in cases of local faults, and, hence, such networks tend to be highly redundant. Their elements are, besides transformers and connections to sources, basically lines and so-called bus-bars acting as nodes in the electrical network. Fig. 2.1 shows a section of a 220/110kV network. Another element of the network is formed by the protection system. Its task is to detect short circuits and respond to them by automatically detaching the affected substructure from the rest of the network by opening switches. Partially conflicting with the goal of optimal protection of the equipment against damage, a guiding principle is to restrict the detached portion to what is really necessary in order to isolate the fault.

2.2 The Models

The models we need for diagnosis have to reflect the behavior of network components under a disturbance of the normal operation (as opposed to modeling the normal situation). While explaining the operation of the protection system, we will
informally describe the component models needed for diagnosis. We exclusively treat only one kind of protection mechanism, called distance protection. Its components are installed between a bus-bar and a line, and (in principle) they act locally. They continuously measure voltage and (direction and magnitude of) current at their location, thus being able to detect the effects of a short-circuit in the network through an increase in current and/or a decrease in voltage. This effect is not limited to the immediate environment of the fault, but may disturb a large portion, or even the entire network. This causes many protections to become active and, thus, the message burst. Hence, the protection system needs a mechanism for deciding which of the activated protections actually have to intervene.

The goal is to detach the smallest substructure necessary to isolate the fault. For instance, line l1 in Fig. 2.3 is protected by breakers br1 and br2, and a fault on bus-bar b1 is isolated by br1, br4 and br5 (this illustrates that distance protections are oriented towards the attached line). To achieve this, each protection determines the distance to the fault location by measuring the impedance at its location. Only if this distance corresponds to the length of the connected line (and energy flows towards it), the protection sends a tripping command to a breaker which disconnects the line. Accordingly, a first model, BRsimpl, for the breaker (including the protection) expresses that it is opened exactly when the fault is on the connected line or the bus-bar behind this line:

\[
\text{STAT} = \text{OPEN} \iff \text{F-DIST} = 1 \lor \text{F-DIST} = 2
\]

where the distance is an abstract one, measured in terms of the number of components to be passed. So, F-DIST = 1 points to the connected line, and F-DIST = 2 to the bus-bar behind it.

The task is then simply to propagate the distance through the network and find the component with F-DIST = 0. The model for correct lines and bus-bars for this purpose simply has to state

\[
\text{F-DIST}_{\text{left}} = 1 \land \text{F-DIST}_{\text{right}} = 1
\]

\[
\land \text{F-DIST}_{\text{left}} + \text{F-DIST}_{\text{right}} = 1,
\]

where F-DIST_{left} and F-DIST_{right} denote the distances at the attached breakers. We will refer to these models of lines and bus bars as L_{DIST} and B_{DIST}, respectively. A model for the only fault, a short circuit, L_{SHORT} and B_{SHORT}, is even simpler:

\[
\text{F-DIST}_{\text{left}} = 1 \land \text{F-DIST}_{\text{right}} = 1.
\]

These simple qualitative models work amazingly well for many standard cases even with incomplete information. However, they are based on a number of assumptions which restrict their applicability. In particular, they may fail if a protection does not work properly. For instance, br2 might not intervene although there is a short circuit on l1. The protection system covers such cases, being more sophisticated than we indicated so far. Actually, the protection distinguishes 4 distance levels, which (slightly simplified) correspond to

- almost the entire length (more precisely, 85% of the impedance) of the directly connected line (for br1 this is l1),
- its rest, the adjacent bus-bar, and part of the lines beyond it (end of l1, b1, and partly l2 and l3), between 85% and 150% of a normalized line impedance,
- the remaining parts of these lines with the bus-bars they are connected to and part of further lines (rest of l2 and l3, b2 and lines beyond it), the interval (150%, 240%)
- everything beyond these three levels, (240%, ∞).

The intervention time is determined to be 0.05s, 0.4s, 0.9s, or 3.0s, dependent on the actual level. Thus, if br2 fails, and, hence, energy flow towards the short circuit continues, br4 and br6 should finally intervene (on level 2 or 3, i.e. after 0.4s or 0.9s). At level 4, the protection also works as a back-up for faults located in the direction of the adjacent bus-bar. A breaker model that reflects the various levels, is the following specialization of BRsimpl, called BRsimpl+:

\[
\text{STAT} = \text{OPEN} \Rightarrow \quad (\text{F-DIST} = 1 \land (\text{LEVEL} = 1 \lor \text{LEVEL} = 2))
\]

\[
\lor (\text{F-DIST} = 2 \land \text{LEVEL} = 2)
\]

But it is still based on the same assumption as BRsimpl, namely that the breaker is not forced to act as a backup for other failing breakers. A model that avoids this simplification is BRdist. Like BRsimpl, it uses the concept of distance in terms of number of components:

\[
\text{STAT} = \text{OPEN} \land \text{LEVEL} = 1 \Rightarrow \text{F-DIST} = 1
\]

\[
\text{STAT} = \text{OPEN} \land \text{LEVEL} = 2
\]

\[
\Rightarrow \text{F-DIST} = 1 \lor \text{F-DIST} = 2 \lor \text{F-DIST} = 3
\]

\[
\text{STAT} = \text{OPEN} \land \text{LEVEL} = 3
\]

\[
\Rightarrow \text{F-DIST} = 3 \lor \text{F-DIST} = 4 \lor \text{F-DIST} = 5
\]

etc.

A more detailed model, BRimpedance, can formulate distances as multiples of the impedance of one line:

\[
\text{STAT} = \text{OPEN} \land \text{LEVEL} = 1 \Rightarrow \text{IMP} < 0.85
\]

\[
\text{STAT} = \text{OPEN} \land \text{LEVEL} = 2 \Rightarrow \text{IMP} < 0.85, 1.5
\]

etc.

We have now obtained a set of models for the correct distance protection which can be organized in a hierarchy as indicated in Fig. 2.4a.
Figure 2.1 Power transmission network

Figure 2.2 Part of a message burst

Figure 2.3 A portion of the network

Figure 2.4 Part of the model graphs a) for breakers and b) for lines
**BR\text{correct}**
is the "ideal" correct model of the behavior of the breaker under normal and disturbed conditions (we may or may not pretend to be able to model it).

**BR\text{impedance}**
is the "strongest" model; it captures all features with high precision.

**BR\text{dist}**
can be obtained from BR\text{impedance} by mapping impedance onto distances counted in number of components.

**BR\text{simpl+}**
is a restriction of BR\text{dist} to F-DIST being 1 or 2, what corresponds roughly to the simplifying assumption that the breaker did not act as a back-up for others.

**BR\text{simpl}**
is like BR\text{simpl+}, but avoids statements about the level.

The interpretation of the edges between the models, in particular the distinction between a strict implication of one model by another one as opposed to one that involves simplifying assumptions, will be discussed in section 5.

Looking at the models for lines and bus bars again, we now have to introduce models that can communicate with BR\text{impedance}. This is no major problem, since we can create, for instance, a model **LIMP-SHORT**:

\[
\text{IMP}_{\text{left}} = z \land \text{IMP}_{\text{right}} = 1 - z \land z \in (0, 1).
\]

However, treating this as the only fault model involves another simplifying assumption: a short circuit does not necessarily imply almost zero resistance; high-resistance faults may occur, for instance, if a broken line touches ground with low conductance. In this case, impedance measured by the protection is higher than for a "standard" short circuit, suggesting a longer distance to the fault. Still, the respective protection could intervene, but at a higher level, and the diagnosis procedure might be misled. Hence, we have to introduce another fault model, **LHIL-RES-SHORT**:

\[
\text{IMP}_{\text{left}} = z + z_{\text{short}} \land \text{IMP}_{\text{right}} = (1 - z) + z_{\text{short}} \land z \in (0, 1) \land z_{\text{short}} \in (0, \infty).
\]

Again, the models of a line (and similarly of a bus) can be organized in a (hyper) graph (Fig. 2.4b):

**LPOSSIBLE**
denotes all physically possible behaviors, correct or faulty.

**LIMPDANCE** and **LDIST**
model this correct behavior, the former in terms of impedance, the latter in terms of F-DIST.

**LIMP-FAULT**
which can be either a choice between LIMP-SHORT and LHIL-RES-SHORT, or consider only the first case under the simplifying assumption that no high-resistance short is present.

**LSHORT**
models the short in terms of F-DIST.

We have obtained graphs of models which become simpler and, hence, less expensive as we approach the leaves. Of course, in parallel, their utility for the diagnostic process might decrease. The strategy for using the graph is to start diagnosis at the (simple and inexpensive) models at the leaves and switch to more expensive models only if necessary. In order to be able to specify this strategy, we have to analyze the various relations between the models and their impact on the diagnoses derivable from them. Some of these relations are based on certain simplifying assumptions, such as "No high-resistance short is present". In the following section, we sketch how we treat such simplifying assumptions as explicit working hypotheses in DP ([Struss 89a]).

3 Diagnosis as a Process
3.1 Model-Based Diagnosis - Where Are We?

The consistency-based approach is currently the mainstream in model-based diagnosis. So far, in systems like GDE ([de Kleer-Williams 87]), Sherlock ([de Kleer-Williams 89]), or GDE* ([Struss-Dressler 89]), and papers like [Reiter 87] and [de Kleer-Mackworth-Reiter 90], the elements of the diagnostic theories were the system description, SD, the observations, OBS, and the possible mode assignments to the elements of COMPS, in the simplest case the choice between CORRECT (or normal) and FAULTY (or abnormal). A diagnosis is then given by a set of faulty components, $\delta \subseteq \text{COMPS}$ such that

\[
\text{SD} \cup \text{OBS} \cup \text{U FAULTY}(C) \cup \text{U CORRECT}(C) \subseteq \delta \subseteq \text{COMPS}
\]

is consistent. Finding diagnoses is strongly driven by exploiting known conflicts. A conflict is a set of mode assignments, $U_{\text{mode}}(C)$, to a number of components, $C_1, ..., C_n \in \text{COMPS}$, that leads to an inconsistency with SD $\cup$ OBS:

\[
\text{SD} \cup \text{OBS} \cup U_{\text{mode}}(C) \vdash \bot,
\]

which is detected mainly through the derivation of contradictory values of one parameter.

The main focus of work in this area and the subject where considerable progress has been achieved concerns the problem: Determine diagnostic candidates given the system description, SD, and the
set of observations, OBS. Formal, sound solutions to this problem have been achieved; however, because of several presumptions and simplifications, they constitute but one element of a theory of diagnosis:

1) They mainly take a static view on diagnosis in that they treat SD as fixed; the structure and potential modifications of SD are rarely discussed. (Exceptions are [Davis 84] and [Preist-Welham 90].) However, during a real diagnostic procedure the system description may be subject to changes. They may concern the structural description (e.g. if the fault is suspected to imply a violation of the original structure), but also a switching to a different behavior model as suggested by the network example.

2) They are not very specific about the nature and control of the overall inference procedure. Most theoretical contributions simply postulate a complete theorem prover, most implemented systems pretend to use a complete constraint propagator. But, as soon as fault models are introduced (if not earlier), complexity problems demand for a focused analysis of the model combinations and a tightly controlled selection of the inferences performed ([Dressler-Farquhar 90], [de Kleer 91]). In practice, diagnostic reasoning is often guided by working hypotheses and simplifications which help reducing complexity but may be revised later, giving rise to non-monotonicties.

3) They do not address the modeling problem but simply assume the existence of powerful and unique models of the device and its components. In contrast, human experts, as we argued in the introduction, may use different, potentially contradictory models dependent on the goal and the stage of the analysis. This includes the exploitation of simplified and approximate models which are known to be ultimately wrong. But we are lacking strong results on automated diagnosis with multiple and simplified models.

The essential characteristic of model-based diagnosis is the exploitation of first principles. Hence, their representation and use has to be the heart of a theory. The first principles are part of SD. As long as all we presume about SD is that it is a set of first order formulas, we can hardly claim to have a theory of diagnosis from first principles.

In summary, what has been consolidated so far, is mainly a theory of determining the space of possible diagnoses and generating diagnostic candidates. Although this is an important step, it is not a sufficient foundation for building diagnostic systems that handle real problems effectively and efficiently.

This purpose requires progress towards a theory of diagnosis that treats diagnosis as a controlled and focused process of acting and non-monotonic reasoning including the use of multiple and simplified models. Hence, rather than facing the task of characterizing the space of possible diagnoses given the conflicts, we try to tackle the problem of how to obtain conflicts, or, more general, information about the consistency of mode assignments, in a controlled and focused manner. In particular, we want to exploit simplifications and to structure the model appropriately. In ([Struss 89a]), we proposed DP as a step in this direction. It has the advantage of generalizing existing formalisms and systems in a very coherent way which even supports an implementation of the generic framework by existing systems, such as GDE (GDE 89).

3.2 DP

Current diagnostic systems are based on a number of assumptions that help to reduce the complexity of the diagnostic task, such as assumptions about the correctness of observations, independence and non-interruption of faults, completeness of knowledge about possible faults, and the system structure being unchanged. These assumptions are all reasonable for many cases, but they limit the applicability of the system in other cases. The problem is that they are present only in an implicit, hardwired form, and, hence, cannot be subject to reasoning and be retracted.

Consequently, DP introduces another element to the theory: the set of diagnostic hypotheses, DHYP, which represent working hypotheses that guide and focus the problem solving process unless they are recognized to be inadequate and dropped. This includes simplifying assumptions and, when working with multiple models, modeling assumptions. As a consequence, our definition of a diagnosis is now the union of a set of faulty components, δCOMP, and a set of diagnostic hypotheses to be retracted, δDHYP.

Definition 3.1 (Diagnosis)

A diagnosis is a set

\[ \delta = \delta_{COMP} \cup \delta_{DHYP} \subseteq \text{COMP} \cup \text{DHYP} \]

such that

\[ \text{SD} \cup \text{OBS} \]

\[ \cup \cup \text{FAULTY}(C) \cup \cup \text{CORRECT}(C) \]

\[ c \in \delta_{COMP} \] \[ c \in \text{COMP} \delta_{COMP} \]

\[ \cup \cup \neg \text{dhp} \cup \cup \text{dhp} \]

\[ \text{dhyp} \in \delta_{DHYP} \]

\[ \text{dhp} \in \text{DHYP} \delta_{DHYP} \]

is consistent.
In other words, we search for a mode assignment to constituents that is consistent with SDUOBS under certain diagnostic assumptions.

At each stage of the diagnostic process the system maintains a set of currently believed diagnostic hypotheses (expectedly a major subset of DHYP) which are not considered for retraction unless the system is forced to, and which, hence, should not appear in a diagnosis. This is expressed by the focus of suspicion. It is given as a subset of the power set of all assumptions,

\[ \text{POS}(\text{COMPS} \cup \text{DHYP}), \]
and describes the set of admissible candidates at a given time. Changing the focus of suspicion is one important diagnostic action of the system. In [Struss 89a] and section 6 of this paper we present examples for this action and its effects. We want to emphasize that, on the one hand, existing consistency-based systems can be regarded as instances of DP. On the other hand, it is easy to realize that assigning TRUE or FALSE to diagnostic hypotheses can be viewed as an analogy to mode assignments to constituents. This provides us with the basis for the implementation of DP, since we can apply an existing diagnostic engine, in our case GDE*, to debug the diagnostic hypotheses as well as the device (see also [Dressler-Struss 92]).

4 Relational Models

We want to analyze the relations between different models of a physical system in order to understand the relations between the sets of diagnoses obtained from using these models. This requires that models are defined in a strict and formal way. Many authors treat models as arbitrary sets of propositions in some logic. This allows us to prove, for instance, whether a certain model contradicts another one. However, such a general characterization does not reflect the nature of the models, namely that they comprise principled knowledge of science and engineering about physical phenomena. This is commonly expressed in terms of (differential) equations over system variables and parameters. Also, the manipulation and transformation of models is naturally done in this descriptive space (think of linear approximation or qualitative abstraction). Logical formulas are usually only a derivate from this primary representation. This is the motivation for our theory of behavior models, model simplification and abstraction based on (mathematical) relations ([Struss 91]). In the following, we will summarize this theory and illustrate its concepts and results by examples from the network domain.

4.1 Models and Behavioral Modes

We take the view that
- a system is composed of some behavioral constituents (such as components or processes),
- the system's behavior is established by the behaviors of its constituents,
- the behavior of some constituent, C, can be specified in terms of local variables \( v_i \in \text{VARS}(C) \), i.e. by a tuple \( \nu_C = (v_1, v_2, ..., v_k) \).

(Completeness: that conceptually, two different constituents do not share variables in common; connectivity is established by stating equality of variable values in SD). If \( \text{DOM}(v_i) \) denotes a possible domain of \( v_i \), then

\[ \text{DOM}(\nu_C) = \text{DOM}(v_1) \times \text{DOM}(v_2) \times ... \times \text{DOM}(v_k) \]

is a space for describing possible behaviors.

Example 4.1

A breaker can be described by its status (i.e. whether it is open or closed), the distance to the short circuit counted in terms of intermediate lines and bus bars, the presence and direction of an over current caused by a short circuit (where a positive sign denotes "towards the line"), and the level of its intervention (corresponding to the distinct delay periods). For convenience, we can combine direction of current and distance, thus creating a directed distance, F-DIST. This establishes a representation

\[ \text{F-DIST} = (\text{STAT}, \text{F-DIST}, \text{LEVEL}) \]

with

\[ \text{DOM}(\text{F-DIST}) = \{\text{OPEN}, \text{CLOSED}\} \times (Z \setminus \{0\}) \times \{0, 1, 2, 3, 4\}, \]

where \( Z \) is the set of integers. Alternatively, we may use impedance instead of the distance (also extended by the sign obtained from the energy flow):

\[ (\text{IMP-IMP}, \text{DOM}(\text{IMP-IMP})) = (\text{STAT}, \text{IMP}, \text{LEVEL}), \]

\[ \{\text{OPEN}, \text{CLOSED}\} \times \mathbb{R} \times \{0, 1, 2, 3, 4\}. \]

Example 4.2

A line can be characterized by the impedance and the duration of energy flow (represented by the level) at its ends:

\[ (\text{XL-IMP}, \text{DOM}(\text{XL-IMP})) = (\text{IMP}_{\text{left}}, \text{IMP}_{\text{right}}, \text{LEV}_{\text{left}}, \text{LEV}_{\text{right}}), \mathbb{R} \times \{0, 1, 2, 3, 4\}^2 \]

Remember, the levels 1 through 4 denote the delay periods 0.0s, 0.5s, 1s, 3s, 5s, respectively, and 0 stands for no over current at all.

The ultimate goal of the diagnostic task is to enable the repair of a system by determining the actual
physical conditions the constituent parts of the system are in, more specifically: whether these conditions allow the normal operation, or which fault is present. For a line, the distinct physical conditions are: an intact wire, a wire connected to ground or another wire with no resistance contact, and a wire with a high-resistance contact to ground or another wire. The diagnostic problem arises because the physical condition is not directly observable (at least not immediately). But it is assumed that each physical condition of a constituent element creates a distinct pattern of behavior and that, hence, the actual physical condition can be determined by identifying the present behavior mode. Stated differently, the physical condition of a constituent restricts the value combinations for its local variables in all real situations we may encounter (under variations in the context this constituent is working in, its exogenous variables, etc.), and a certain mode of behavior can be described by this restricted subset, i.e. a relation $R \subseteq \text{DOM}(y_R)$.

**Example 4.3**

The relation $R_{\text{BR-IMP}} \subseteq \text{DOM}(y_{\text{BR-IMP}})$ describing the correct behavior of a breaker in terms of $y_{\text{BR-IMP}} = (\text{STAT}, \text{IMP}, \text{LEVEL})$ is extensionally listed in Table 4.1a and shown in Fig. 4.1a. In the cases where the protection does not trip the breaker, the level should again be interpreted at the duration of the disturbance. (Note that we ignore durations of short circuits other the fixed time thresholds, e.g. due to the short vanishing for different reasons. A temporal model would complicate the examples too much).

**Example 4.4**

The three behavior models of the line can be described by relations as follows: The relation $R_{\text{L-IMP}}$ (Table 4.2) states that the (positive) impedance observed is not less than the impedance of a line, which is normalized to 1. At the other end its magnitude is reduced by 1 and the sign changed. Finally, the duration of the overcurrent at both ends must be the same. The projection of the relation into the $\text{IMP}_{\text{left}}-\text{IMP}_{\text{right}}$ plane is shown in Fig. 4.2a.

$R_{\text{L-IMP-SHORT}}$ (Table 4.3) expresses that breakers at both ends observe impedances less than 1 and positive (i.e. pointing towards the line). Note that, in contrast to the correct line, the durations of the disturbance at the ends are not related, because they depend on the reaction of the breakers on the respective side, which may be different. It covers the "asymmetric case", which may occur, for instance, when only one of the ends of a broken wire is in contact with the ground.

Finally, $R_{\text{HI-RES-SHORT}}$ (Table 4.4) describes the high-resistance short: again (except for the no-over-current case and the asymmetric fault) the signs of the impedance are positive at both ends, but in this case, they add up to a value greater than 1, due to the non-zero resistance at the shorted part (Fig. 4.2c).

How can these relational descriptions of behavior modes be used to achieve the goal of identifying the actual behavior mode (and, through this, the present physical condition)? We can observe a set of real situations and check whether or not the observations are covered by the relations.

For fundamental reasons we are never able to positively confirm a particular behavior mode, unless we make further assumptions. Even if $\text{DOM}(y_R)$ or the relation $R$ is finite and we observed all elements of $R$ in reality and no value contradicting $R$, we cannot conclude that future situations may reveal observations not be covered by $R$. But what is possible and fundamental for consistency-based diagnosis is that we are able to refute a behavior mode if we encounter a situation with values outside the respective relation. For instance, if we measure an impedance less than 1 (normal line impedance) with direction towards a line (i.e. with sign "+"). $R_{\text{L-IMP}}$ does not cover this situation, and the correct mode of the line is refuted.

So the basic rule of the game in consistency-based diagnosis is to take a set of relations describing the behavior modes we are aware of or interested in, expose the respective models to the rough climate of real situations and observe which ones survive. This will work as long as the models have a certain quality: the respective relation has to be guaranteed to cover all situations permitted by the behavioral mode it describes. It must not be smaller than the "real" one. But it may be larger! This is important, because it helps to cope with a fundamental problem of model-based reasoning: we may lack an ideal model of the real behavior mode, or, stated more aggressively, we never have an exact model of a physical phenomenon. But we are not forced to have it in order to perform consistency-based diagnosis: The odd-shaped relation describing a certain behavior mode in Fig. 4.3 may be unknown to us; but if we can be sure the relation $R$ covers this unknown relation, it may help. From observing or inferring a value $y_0$ in $\text{DOM}(y_R)$ for which $y_0 \in R$ holds, we can conclude that $y_0$ is not in the unknown "ideal" relation and, hence that the respective mode is not present.
Table 4.1a R_{BR-IMP}, the correct behavior of a breaker. It is the union of the lines of the table; for instance, the first line is to be read (OPEN) × (0, 1.5) × (1).

Table 4.1b R_{BR-DIST} = t_{DIST}(R_{BR-IMP})

Table 4.1c R_{SIMPL+}, describing the breaker in the cases where it does not work as a back-up

Table 4.1d The difference between the distance model and the simplified one
Table 4.2  $R_{L-Imp}$, describing the correct line

<table>
<thead>
<tr>
<th>IMP$_{left}$</th>
<th>IMP$_{right}$</th>
<th>LEV$_{left}$</th>
<th>LEV$_{right}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>($\rightarrow$, 0)</td>
<td>1-IMP$_{left}$</td>
<td>${1, 2, 3, 4}$</td>
<td>=LEV$_{left}$</td>
</tr>
<tr>
<td>(1, $\rightarrow$)</td>
<td>1-IMP$_{left}$</td>
<td>${1, 2, 3, 4}$</td>
<td>=LEV$_{left}$</td>
</tr>
</tbody>
</table>

Table 4.3  $R_{L-Imp-Short}$, the normal short

<table>
<thead>
<tr>
<th>IMP$_{left}$</th>
<th>IMP$_{right}$</th>
<th>LEV$_{left}$</th>
<th>LEV$_{right}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0, 1)</td>
<td>0</td>
<td>${1, 2, 3, 4}$</td>
<td>0</td>
</tr>
<tr>
<td>(0, 1)</td>
<td>1-IMP$_{left}$</td>
<td>${1, 2, 3, 4}$</td>
<td>${1, 2, 3, 4}$</td>
</tr>
</tbody>
</table>

Table 4.4  $R_{Hi-Res-Short}$, high-resistance short

<table>
<thead>
<tr>
<th>IMP$_{left}$</th>
<th>IMP$_{right}$</th>
<th>LEV$_{left}$</th>
<th>LEV$_{right}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0, $\rightarrow$)</td>
<td>0</td>
<td>${1, 2, 3, 4}$</td>
<td>0</td>
</tr>
<tr>
<td>(0, $\rightarrow$)</td>
<td>(0, $\rightarrow$)</td>
<td>${1, 2, 3, 4}$</td>
<td>${1, 2, 3, 4}$</td>
</tr>
<tr>
<td>IMP$<em>{left}+$IMP$</em>{right}&gt;1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.5  $R'_{Hi-Res-Short}$, describing the "ideal" high-resistance short

<table>
<thead>
<tr>
<th>IMP$_{left}$</th>
<th>IMP$_{right}$</th>
<th>LEV$_{left}$</th>
<th>LEV$_{right}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0, $R_{max}$)</td>
<td>0</td>
<td>${1, 2, 3, 4}$</td>
<td>0</td>
</tr>
<tr>
<td>(0, $R_{max}$)</td>
<td>(0, $R_{max}$)</td>
<td>${1, 2, 3, 4}$</td>
<td>${1, 2, 3, 4}$</td>
</tr>
<tr>
<td>IMP$<em>{left}+$IMP$</em>{right}&gt;1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example 4.5

If we consider RHRES-SHORT, the relation describing the high-resistance short on a line, we have to state that it is definitely not exactly describing this kind of faults, as the following consideration indicates. The difference between the sum of IMP_left and IMP_right and 1 is due to the resistance of the contact. What is expressed in RHRES-SHORT about this resistance is only that it is positive. Let us ignore the problem of how to distinguish a small value from the “ideal” short circuit. But as we let the resistance go to infinity, there will be a threshold beyond which the short will no longer cause an over current which triggers the protection system because it is in the order of magnitude of the current under normal conditions.

Remember that the positive sign of the impedance (and of the distance, likewise) expresses the direction and, hence, the existence of a current greater than a certain threshold. As a result, a large subset of RHRES-SHORT does not correspond to a really possible situation, and in reality the relation describing this case, RHRES-SHORT (Table 4.5), is bounded as indicated in Fig. 4.2d.

However, although we know there exist upper bounds, R_max, for the impedance, we may not be able to determine it exactly. The principal reason is not only imprecision of measurements or insufficient information about the particular line and its adjacent breakers, but the fact that R_max depends on the global context given by the current network topology and state of the network. But we may use RHRES-SHORT as an appropriate substitute; if it is rejected by observations (e.g. IMP_left<0), then so is the “ideal” model, and, hence, we can safely rule out a high-resistance short.

We hope to have motivated the following definition of a relational model which reflects the principle of consistency-based diagnosis discussed above.

Intuitively speaking, a model is the claim (or the guess) that a certain relation, R, covers what is permitted by the current physical condition in any situation, s∈SIT, we might encounter, i.e. that R covers a particular behavioral mode (Fig. 4.3).

The set of situations which are physically possible due to the actual physical condition of a constituent and, perhaps, due to environmental conditions will be denoted SIT. If, in a particular situation, s∈SIT, y_0 has the value y_0∈DOM(y_0), we will write

Val(s, y_0).

Figure 4.3 A model covers the physically possible values

Definition 4.1 (Model)

A relation R⊆DOM(y_0) specifies a model of a constituent C by

M(C, R) ⇔

∀y_0∈DOM(y_0) \ ( (∃s∈SIT Val(s, y_0, y_0)) ⇒ y_0 ⊆ R ).

Observation of a value, y_0, in a situation, s, implies that the Val-predicate holds:

Obs(s, y_0) ⇒ Val(s, y_0).

Then Definition 4.1 suffices to realize the following inference:

If OBS ⊆ SD ⊆ Val(s, y_0)

then OBS ⊆ SD ⊆ M(C, R).

Note that in our theory a model is not merely a “descriptive term”, such as an equation or a differential equation. A model is rather the result of an interpretation of such a descriptive term. For instance, the equation

v/I = R_0

is not the model of a resistor, but it can be the origin of a model (or of several). This could be the statement that the possible values of voltage and current are covered by the relation

{(v, i)∈ R^2 \mid v/i = R_0}.

The reason for this approach is not only the fundamental argument that, unless we specify an interpretation, the equation is merely a string which cannot be related to any measured or inferred value. In fact, there can be several useful interpretations, leading to different models. Other models of the resistor could be specified by the relations

{(v, i)∈ R^2 \mid v/i ∈ (.9 R_0, 1.1 R_0)}

(taking into account changes in the resistance and/or imprecision of measurements, see [Struss 91]), or

{(+, +), (0, 0), (-,-)},

which interprets the equation in the domain of signs.
If we are (or pretend to be) sure that a particular relation describes a behavior mode exactly, this means it does not only cover it, but also each of its elements can be realized in a situation. This is captured by the following definition:

**Definition 4.2 (Strong Model)**

A relation $R \subseteq \text{DOM}(\gamma_c)$ specifies a strong model of $C$ by

$$B(C, R) \iff \forall y_0 \in \text{DOM}(\gamma_c) \exists s \in \text{SIT} \text{ Val}(s, \gamma_c, y_0) \Rightarrow y_0 \in R.$$  

Strong models can be used as representatives of the real behavioral modes. Thus, even if we are unable to explicitly describe the respective relation, we can integrate the behavioral modes in the model relation graph.

For each behavior mode, $m$, to be considered, we introduce (exactly) one designated strong model $B_m(C, R)$, potentially with $R$ unspecified, in the graph. Any model $M(C, R')$, which is entailed by it,

$$B_m(C, R) \Rightarrow M(C, R'),$$

will be called a model of the mode $m$. For instance,

$$R \subseteq R' \subseteq \text{DOM}(\gamma_c)$$

is a sufficient condition for this situation, $\text{lCorrect}$ in the model graph of the line is such a behavioral mode.

If $M(C, R)$ is consistent with the currently possible situations (or, rather, with the information obtained about them, so far) and, hence, can be a model of the behavior mode, we call it valid model. Obviously, an important problem we have to analyze is which changes in modes preserve the validity, perhaps under certain additional assumptions. This is the topic of the next sub-section.

Strong models (and, hence, behavior modes) are exclusive if specified over the same domain. Because they claim to describe exactly what is possible, only one of them can be right if they differ in one tuple of values:

$$R, R' \subseteq \text{DOM}(\gamma_c) \land R \neq R' \Rightarrow (B(C, R) \Rightarrow \neg B(C, R')).$$

This does not exclude overlap of the respective relations. For instance, all behavior modes of the line share the tuple $(0, 0, 0, 0)$. Two relations are equal if and only if they are equal as sets. A mode can even be specified by a superset of the relation of another mode, for example, for an intermittent behavior (see [Struss 91]). Secondly, exclusiveness of strong models does not preclude the existence of several valid strong models, given information about a certain set of situations, which will always be limited.

Note, that "exactness" of strong models is a relative one w.r.t. the precision provided by the representation $(\gamma_c, \text{DOM}(\gamma_c))$. For instance, we might choose

$$\text{DOM}(\gamma_{L-\text{IMP}}) = \{-, 0, +\} \times \{1, 2, 3, 4\}$$

by taking the signs of the impedance. Then $R_{L-\text{IMP}}(\text{SIGN})$ (Table 4.6) specifies a strong model of both the normal short and the high-resistance short. It might appear as a clear disadvantage that the model based on $R_{L-\text{IMP-SIGN}}$ is unable to discriminate the two fault modes. However, this can be a benefit, because a refutation of this single model based on observations or inferences suffices to eliminate the two faults without having to instantiate other models. Again, this example illustrates the value of multiple models.

<table>
<thead>
<tr>
<th>$R_{L-\text{IMP-SIGN}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP$_{\text{left}}$</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>+</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>+</td>
</tr>
</tbody>
</table>

Table 4.6 $R_{L-\text{IMP-SIGN}}$, covering all types of shorts

[Friedrich-Gottlob-Nejdl 90] suggests that we might also want to state what is impossible:

**Definition 4.3 (Impossibility)**

A relation $R \subseteq \text{DOM}(\gamma_c)$ defines a (weak) impossibility, by

$$I(C, R) \iff \forall y_0 \in \text{DOM}(\gamma_c) \exists s \in \text{SIT} \text{ Val}(s, \gamma_c, y_0) \Rightarrow y_0 \notin R.$$  

**Example 4.6**

If $R_{L-1} = \{\text{IMP}_{\text{left}}, \text{IMP}_{\text{right}}, \text{LEV}_{\text{left}}, \text{LEV}_{\text{right}}\}$, $0 < \text{IMP}_{\text{left}}/\text{IMP}_{\text{right}} < \eta$, then $I(L, R_{L-1})$ is a valid impossibility if the correct mode is present, but not in the case of a short.

Note, however, that impossibility does not contain different information, since we have the obvious equivalence

$$I(C, R) \iff M(C, \text{DOM}(\gamma_c) \neg R)$$

(Fig. 4.4a). Also there is no technical difference between the two ways of representing knowledge about the constituent; both can be used for prediction as well as for mere consistency checking. For the latter purpose, those impossibilities are useful that hold always, i. e. independently of the physical condition of the constituent (Fig. 4.4b).
Definition 4.4 (Physical Impossibility)
\( M(C, R) \) is called a physical impossibility iff
\( M(C, \text{DOM}(\gamma_C))R \)
is a model of all physically possible behavioral modes of \( C \).

![Figure 4.4 a) Impossibility](image)

Definition 4.5 (Representation)
A representational space, for short: representation, is a pair
\( (\gamma_C, \text{DOM}(\gamma_C)) \).

A model is specified by some relation \( R \subseteq \text{DOM}(\gamma_C) \).
Hence, a change in a model may be caused by the use of
- different variables, \( \gamma_C \), or
- different domains, \( \text{DOM}(\gamma_C) \), for the same variables, or
- different relations for defining a model while \( \gamma_C \) and \( \text{DOM}(\gamma_C) \) remain fixed.
The first two cases imply a shift in representation, whereas the third preserves it and directly modifies the model.

Example 4.8
The step from \( \text{BRIMPEDANCE} \) to \( \text{BRDIST} \), i.e. from
\( \gamma_{\text{BRIMP}} = (\text{STAT}, \text{IMP}, \text{LEVEL}) \)
to
\( \gamma_{\text{BRDIST}} = (\text{STAT}, \text{F-DIST}, \text{LEVEL}) \),
is an example for the first case.

Example 4.9
A transformation that has to be applied (implicitly) before the one of Example 4.6 maps current from a real-valued domain to three symbolic values characterizing the range below a certain threshold (NORM) and an over current (OC) in either direction:
\( R \rightarrow \{\text{OC}, \text{NORM}, \text{OC}'\} \).
This illustrates the second case.

Example 4.10
An instance of the third case is the modification of \( \text{BRDIST} \) which lead to \( \text{BRSIMPL} + \) by changing the status from OPEN to CLOSED for
\( \text{F-DIST} > 2 \) or \( \text{F-DIST} < 0 \).
Let us first look at cases 1 and 2, i.e. changes in \( (\gamma_C, \text{DOM}(\gamma_C)) \). A "reasonable" transformation that turns one representation into another one should respect the Val predicate (introduced in section 4.1) in the following sense:
- **Val-preserving**: If Val holds for a value in the original representation, then it also holds for the transformed value. For instance: if we map real numbers to \( \{\text{OC}, \text{NORM}, \text{OC}'\} \), as in Example 4.9, then \( \text{Val}(s, \text{iGR}, 220000) \) implies \( \text{Val}(s, \text{iGR}, \text{NORM}) \).
- **Val-grounding**: If Val holds for a transformed value, this must be the case for one of its preimages in the original representation: \( \text{Val}(s, \text{iGR}, \text{NORM}) \) is true only if there exists some positive real number, \( r \), below the threshold for over current such that \( \text{Val}(s, \text{iGR}, r) \).
We will call it a **representational transformation**, if it is Val-preserving and Val-grounding. The fundamental theorem for representational...
transformations (proof in [Struss 91]) is the following:

Theorem 4.1
Let \( R \subseteq \text{DOM}(y_0), R' \subseteq \text{DOM}(y_1) \), and
\( \tau_0 : \text{DOM}(y_0) \to \text{DOM}(y_1) \).

If \( \tau_0 \) is Val-grounding then the image of a valid model is a valid model:
\[ M(C, R) \Rightarrow M(C, \tau_0(R)). \]

If \( \tau_0 \) is Val-preserving, then the inverse transformation also preserves the validity of models:
\[ M(C, R) \Rightarrow M(C, \tau_0^{-1}(R')). \]

If \( \tau_0 \) is a representational transformation, then the image of a valid strong model is a valid strong model:
\[ B(C, R) \Rightarrow B(C, \tau_0(R)). \]

To give a precise meaning to a widely used term, we consider abstraction to be a representational transformation that is surjective (i.e. each element of \( \text{DOM}(y_1) \) is the image of an element of \( \text{DOM}(y_0) \)) and is not injective (i.e. some elements of \( \text{DOM}(y_1) \) are the image of more than one element of \( \text{DOM}(y_0) \)). Then Theorem 4.1 implies that abstraction and its inverse preserve validity of models.

Example 4.11
Considering again the shift from BR\textsubscript{IMPEDANCE} to BR\textsubscript{DIST}, we can define a transformation
\[ \tau_{\text{DIST}} : \text{DOM}(y_{\text{BR-IMP}}) \to \text{DOM}(y_{\text{BR-DIST}}) \]
\[ \tau_{\text{DIST}} : (\text{STAT, IMP, LEVEL}) \to (\text{STAT, F-DIST, LEVEL}) \]
by \( \tau_{\text{DIST}} = (\text{id}_{\text{STAT}}, \tau_{\text{INT}}, \text{id}_{\text{LEVEL}}) \), which leaves status and level identical and maps reals representing multiples of the normal line impedance to integers counting the number of components by
\[ \tau_{\text{INT}} : R \to Z \setminus \{0\} \]
with
\[ \tau_{\text{INT}}(r) = \begin{cases} 2[r] + 1 \text{ if } r > 0 \\ 2[r] - [r] + 1 \text{ if } r \leq 0 \end{cases} \]
where \([r]\) is the smallest integer greater than or equal to \(r\).

Due to its semantics, \( \tau_{\text{DIST}} \) is a representational transformation, and Theorem 4.1 tells us that
\[ M(BR, R_{BR-IMP}) \Rightarrow M(BR, \tau_{\text{DIST}}(R_{BR-IMP})). \]
The relation obtained, \( \tau_{\text{DIST}}(R_{BR-IMP}) \) is in fact \( R_{BR-DIST} \) (Table 4.1b), the one defining model BR\textsubscript{DIST}. It is shown in Fig. 4.1b. Because \( \tau_{\text{DIST}} \) is surjective (i.e. completely covers \( Z \{0\} \)) and not injective (mapping all reals on one line onto one integer), represents an important class of abstraction transformations including interval-based qualitative abstraction ([Struss 88b, 89b]).

Example 4.12
The transformation of BR\textsubscript{SIMPL}+ that creates BR\textsubscript{SIMPL} by eliminating the variable LEVEL, i.e. the projection
\[ \tau_{\text{LEVEL}} : (\text{STAT, F-DIST, LEVEL}) \to (\text{STAT, F-DIST}) \]
is another example of an abstraction. It yields
\[ BR_{SIMPL}+ \to BR_{SIMPL}. \]

Under abstraction, discriminating power of models (w.r.t. behavioral modes) may be lost, and under the inverse mapping, a strong model will often lose its strength; but still, the images of valid models are valid models. This can also be the case if the representation remains the same, but the specifying relation is changed, which was the third case in our list of changes. Some possible changes are covered by the following lemma.

Lemma 4.2
\[ R \subseteq R' \subseteq \text{DOM}(y_0) \Rightarrow (M(C, R) \Rightarrow M(C, R')) \]
\[ (M(C, R) \cup M(C, R') \Rightarrow M(C, R \cup R')) \]
\[ (M(C, R) \land M(C, R') \Rightarrow M(C, R \cap R')) \]

Example 4.13
In the power transmission network, the component we called breaker is actually an aggregate comprising the distance protection, DPr, and a switch (the real breaker), SW. The former measures current and voltage, estimates the distance to the fault based on the impedance, and, if the conditions discussed earlier are satisfied, sends a tripping command to the switch which then breaks the connection. Figure 4.5 shows the refined model revealing in addition to
\[ y_{BR-IMP} = (\text{STAT, IMP, LEVEL}) \]
another variable which was hidden before, the command, CMD.

![Figure 4.5 The expanded model of the breaker](image)

If we ignore the line transmitting this command, the representations for the subcomponents are
(\forall \psi, \text{DOM}(\psi)) =
(\text{CMD, IMP, LEVEL}),
\{\text{TRIP, STAT}\} \times \mathbb{R} \times \{0, 1, 2, 3, 4\},
(\forall \varphi, \text{DOM}(\varphi)) =
(\text{CMD, STAT}),
\{\text{TRIP, STAT}\} \times \{\text{OPEN, CLOSED}\},
\text{respectively. The modeling relation for the correct behavior of the distance protection, } R_{DP-IMP}, \text{ is obtained from } R_{BR-IMP} \text{ (Table 4.1a and Fig. 4.1a) by replacing OPEN and CLOSED by TRIP and STAT, respectively. The correctly functioning switch is specified by}
\begin{align*}
R_{SW} &= \{\text{TRIP, OPEN}, \text{STAT, CLOSED}\}.
\end{align*}
\text{We embed the relations, } R_{DP-IMP} \text{ and } R_{SW}, \text{ in an extended representation for the breaker:}
(\forall \varphi, \text{DOM}(\varphi)) =
(\text{STAT, CMD, IMP, LEVEL}),
\{\text{OPEN, CLOSED}\} \times \{\text{TRIP, STAT}\} \times \mathbb{R} \times \{0, 1, 2, 3, 4\},
\text{by mappings } e_1, e_2. \text{ Then the intersection of the embedded relations,}
R_{BR-IMP+} \equiv e_1(R_{DP-IMP}) \cap e_2(R_{SW}),
\text{describes again the correct behavior of the whole breaker, BR.}
\text{In our theory, we are able to prove that the behavior model of the compound system is an implication of the models of its parts. The embeddings are representational transformations, and Theorem 4.1 gives us}
M(DPR, R_{DP-IMP}) \Rightarrow M(DPR, e_1(R_{DP-IMP})),
M(SW, R_{SW}) \Rightarrow M(SW, e_2(R_{SW})).
\text{With Lemma 4.2, we obtain}
M(DPR, e_1(R_{DP-IMP})) \land M(SW, e_2(R_{SW})),
\Rightarrow M(BR, R_{BR-IMP+}).
\text{The projection of } R_{BR-IMP+} \text{ to } \text{DOM}(R_{BR-IMP}), \text{ which drops the variable CMD, regains } R_{BR-IMP}, \text{ the modeling relation used so far. This mapping is a representational transformation, too, and, hence,}
M(BR, R_{BR-IMP+}) \Rightarrow M(BR, R_{BR-IMP})
\text{holds. The overall result is}
M(DPR, R_{DP-IMP}) \land M(SW, R_{SW}),
\Rightarrow M(BR, R_{BR-IMP}),
\text{which states that the behavior model of the entire breaker is implied by the models of its parts. This illustrates that the relational formalism and, in particular, the concept of representational transformations provide a sound basis for handling hierarchical models and structural refinement.}
\text{If we apply some arbitrary "surgery" directly to the model, we cannot guarantee that a valid model is turned into another valid model. Simplifications can be regarded as mapping the relations specifying models onto different relations in the same representational space.}
\text{Example 4.14}
\text{The shift from } BR_{DIST}, \text{ which covers all levels, to } BR_{SIMPL+} \text{ provides an example for this case. We obtained the relation specifying } BR_{SIMPL+} \text{ from } BR_{DIST} \text{ basically by changing the tuples corresponding to } STAT=OPEN \text{ with negative distances or distances greater than } 2 \text{ (Table 4.1c). Fig. 4.1c shows that, in comparison to Fig. 4.1b, the relation has been significantly changed.}
\text{We cannot expect to obtain any strict relationship of the models specified by these two relations. For instance, observing } \forall \psi, \text{DIST}_0=\{\text{OPEN, 3, 2}\} \text{ would invalidate } M(BR, R_{SIMPL+}), \text{ but not } M(BR, R_{DIST}). \text{ On the other hand, } \forall \psi, \text{DIST}_0=\{\text{CLOSED, 3, 4}\} \text{ contradicts the latter but not the former model. } M(BR, R_{DIST}) \text{ specifies a model of the correct breaker. If we modify the relation, we run the risk of refuting the correct mode based on observation (or inference) of a value that corresponds to the real behavior but is not covered by simplified relation, e.g. } \forall \psi, \text{DIST}_0=\{\text{OPEN, 3, 2}\}. \text{ This is not a risk as long as we assume that situations with such values do not occur. This means we have to assume that}
\text{Val}(\zeta, \psi, \varphi) \Rightarrow \forall \psi \in \text{CORRECT} \Rightarrow \text{SIMPL+}
\text{where } \text{CORRECT} \text{ is the ideal relation describing the mode. This may be unknown. But } R_{DIST} \text{ is considered to cover } \text{CORRECT}, \text{ and thus, we are able to state a sufficient condition for a simplified relation to establish a valid model. This is expressed in the following lemma. Intuitively, this lemma simply states that the modified relation still defines a model, if the parts of } R \text{ it truncates do not correspond to encountered situations. (Fig. 4.6).}
\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure4.6.png}
\caption{M(C, R') as a simplification of M(C, R)}
\end{figure}
Lemma 4.3
Let \( R, R' \subseteq \text{DOM}(\gamma) \).
\[
(\forall \gamma_0 \in \text{DOM}(\gamma))
\quad (\exists s \in \text{SIT} \quad \text{Val}(s, \gamma_0, \gamma_0) \Rightarrow \gamma_0 \in \text{RR}(R'))
\Rightarrow (M(C, R) \Rightarrow M(C, R')).
\]
The condition of lemma 4.3 can be seen as the assumption that the physical condition of \( C \) and/or environmental conditions restrict the possible physical situations accordingly.

Example 4.14 (Cont'd)
To apply Lemma 4.3 to the simplification of the breaker model, we have to determine the difference between the respective relations (Table 4.1d). Treating \( \text{BR}_{\text{SIMPL}} \) as a model for the correct breaker is justified, if no situations with these values are encountered. The following formula is a sufficient condition:
\[
\forall s \in \text{SIT} \quad \neg \text{Val}(s, \text{SIT}_{\text{DIST}}(\text{OPEN}, 3, 2))
\quad \land \quad \neg (\text{Val}(s, \text{LEVEL}, \text{lev}) \land \text{lev} \geq 3)
\]
In reality, it is satisfied, if the disturbance is not a high-resistance fault, if breakers closer to the fault intervene properly, and if the breaker itself responds to any fault on the adjacent line and the bus-bar behind it. This reproduces exactly the original motivation for this model.

If we change a model by manipulating the specifying relation, then, in general, the model obtained is not a strict implication of the original one, but rather a conditional implication:
\[
M(C, R) \land \text{ASSUMPTION} \Rightarrow M(C, R'),
\]
where ASSUMPTION may impose some restrictions on \( R, R' \) and the situations encountered, as in Lemma 4.3. This has an important impact on the results of consistency-based diagnosis using simplified models which is discussed in the next section.

We conclude this subsection by summarizing that we consider
- abstraction as a special transformation of representations (that induces a transformation of models)
- and simplification as a transformation of models within the same representation.

The former preserves model properties while the latter may violate them, but in a way that can be described. Approximations are special cases of simplifications that allow measuring the proximity of models.

5 Model Relations
In the preceding section, we emphasized the distinction between a strict implication and a conditional implication of models. We did so because switching between models has a different impact on the resulting diagnoses dependent on the type of the link. Of course, other links between models can be imagined and are indeed used in existing systems, explicitly or implicitly. The step from a black box model of an aggregate to its expanded model is one example. To study the impact of model switching on the results of consistency-based diagnosis, it is necessary and sufficient to consider the logical relations between the models rather than the specific transformation that created this logical relation. This simplifies our analysis because there are fewer logical links between models than motivations and techniques for manipulating models. This is why the theory of consistency-based diagnosis with multiple models we develop in the following section, is independent of the particular modeling formalism (and, hence, we simply denote models by \( M, M' \), etc.). It only requires that the models can be seen as propositions and that logical connectors can be established between them. Our relational formalism is one instance and will be used to provide examples.

Consistency-based diagnosis is driven by detecting discrepancies based on observing and predicting values of variables. Hence, there must exist a set of explicit models which can be used directly for prediction of constituent behavior, e.g. relational models defined by some (extensionally or intensionally) specified relation.

Besides such explicit models, we would also like to represent concepts that have no predictive power directly associated with them. For example, the ideal correct behavior mode of a breaker, \( \text{BR}_{\text{CORRECT}} \), may not be described explicitly; yet it is possible to relate it to abstract or simplified models of this mode, as was exemplified in the previous section. A concept like \( \text{BR}_{\text{FAULTY}} \) is also not an explicit model, but rather specified as a choice between different possible fault modes which might be represented by explicit models. In order to include such "implicit models", we generalize the concept of a model by the following recursive definition.

Definition 5.1 (Generalized Model)
A (generalized) model is
- an explicit model, or
- a disjunction of (generalized) models, or
- a conjunction of (generalized) models, or
- the antecedent of a view or of a simplification (to be defined below) that is a (generalized) model.

5.1 View
From a theoretical point of view, the best way to model a device is to construct and use strong models
only and to give a precise description of all aspects of behavior. However, besides the fundamental objection that our knowledge may not suffice, reasoning with strong models is often too expensive and in many cases unnecessary, as we argued before. This is why we want to split the model and maintain models at different levels of abstraction and simplification. Sometimes, working with a weaker model is much less expensive, but still effective.

Revisiting the models of the network components (Fig. 2.4) and considering the analysis in the previous section, we notice that some links connect a model with a strictly weaker version of itself. For instance, $\text{BR}_\text{SIMPL+}$ imposes the same constraints on $\text{P-DIST}$ as $\text{BR}_\text{SIMPL+}$, but no constraints on $\text{LLEVEL}$; it allows less inferences. Hence, we have a strict logical implication:

$$\text{BR}_\text{SIMPL+} \Rightarrow \text{BR}_\text{SIMPL}.$$ 

In consistency-based diagnosis, the contraposition is more important, because from a refutation of $\text{BR}_\text{SIMPL+}$ we can conclude the refutation of $\text{BR}_\text{SIMPL+}$ without having to instantiate this model:

$$\neg \text{BR}_\text{SIMPL} \Rightarrow \neg \text{BR}_\text{SIMPL+}.$$ 

In the following, we will call such a related model a view:

**Definition 5.2 (View)**

A model $M'$ is a view of another model, $M$ iff $M \Rightarrow M'$.

In other words, $M'$ is a necessary condition for $M$ to hold.

**Example 5.1**

The simplest example is that a superset $R' \supseteq R$ defines a view $M(C,R')$ of $M(C,R)$ (Fig. 5.1):

$$R \subseteq R' \Rightarrow (M(C,R) \Rightarrow M(C,R')).$$

![Figure 5.1 M' is a view of M](image)

Consider again the modeling of the high-resistance short discussed in Examples 4.4 and 4.5. If we are unable to exactly determine the threshold $R_{\text{max}}$ beyond which a high-resistance short does not create an over current that triggers the protection system, we can use some substitute for $R_{\text{max}}$ that is definitely greater, e.g. $R_{\text{max}} = \infty$. This corresponds to replacing the unknown ideal relation by the strictly larger one (Figures 4.2c,d). In our terminology

$$L_{\text{HI-RES-SHORT}} = M(L, R_{\text{HI-RES-SHORT}})$$

is a view of

$$L_{\text{HI-RES-SHORT}} = M(L, R_{\text{HI-RES-SHORT}}'),$$

and can be used to rule out this fault mode.

**Example 5.2**

From Theorem 4.1, it follows directly that representational transformations in general and model abstraction in particular lead to a view: If $M(C, R')$ is an abstraction of $M(C, R)$ then $(M(C, R) \Rightarrow M(C, R'))$.

In example 4.11, we discussed the mapping of intervals of real numbers (representing values of impedance) to integers (distance as number of components). This is a representational transformation and, more specifically, an abstraction that generates the relation specifying the breaker model

$$\text{BR}_\text{DIST} = M(\text{BR}, R_{\text{BR-DIST}})$$

(Fig. 4.1b) as the image of the relation of the impedance model,

$$\text{BR}_\text{IMP} = M(\text{BR}, R_{\text{IMP}}).$$

Again,

$$\text{BR}_\text{IMP} \Rightarrow \text{BR}_\text{DIST}$$

holds which allows us to use $\text{BR}_\text{DIST}$ as a necessary condition for $\text{BR}_\text{IMP}$. 

**Example 5.3**

The original motivation for introducing views was the idea to distinguish different (physical) aspects of behavior, to reason separately about them, and consider only as few as necessary. To illustrate this idea, consider again the impedance model of the correct line, given by the relation $R_{\text{L-IMP}}$ (see Table 4.2). It captures two different aspects: First, there is no short circuit on the line causing an over current (this is expressed by the restriction on the impedances). Second, if there is an over current (caused by some other component), then its duration is the same at both ends of the line (expressed by equality of the levels). We can (almost) split $R_{\text{L-IMP}}$ into two orthogonal relations:

$$R_{\text{L-IMP}} = R_{\text{L-IMP-Z}} \times R_{\text{L-IMP-LV}} \cup \{(0, 0, 0, 0)\},$$

where

$$R_{\text{L-IMP-Z}} = \{(\text{IMP}_{\text{left}}, \text{IMP}_{\text{right}}) \mid \text{IMP}_{\text{left}} \oplus \text{IMP}_{\text{right}} = 1 \land \text{IMP}_{\text{left}} \in [0, 1] \} \subseteq R^2,$$
RL-IMP-LV =
((LEV_{left} \cup \text{LEVright}) \mid
\text{LEV}_{left} = \text{LEV}_{right}) \subseteq N^2.

Note that (0, 0, 0, 0), the case of no or normal current, is the only link between the two aspects. Both partial relations, with (0, 0) added, can be regarded as projections (and, hence, abstractions) of RL-IMP, and we have gained two views of IMP:

\[ M(L, RL-IMP) \Rightarrow M(L, RL-IMP \cup \{(0,0)\}) \]
\[ M(L, RL-IMP) \Rightarrow M(L, RL-IMP-LV \cup \{(0,0)\}) \]
which can be used independently. Actually, this can be of practical interest, because in some network control systems, information about the intervention level is not available, and temporal information is not reliable. In this case, it makes sense to use \( M(L, RL-IMP-LV) \) instead of \( L_{IMP} \), and, if necessary, consider additionally \( M(L, RL-IMP-LV) \) based on hypothesize-and-test for determining the level.

Trivially, the view relation is transitive:

Remark 5.1
If \( M' \) is a view of \( M \), and \( M'' \) is a view of \( M' \), then \( M'' \) is a view of \( M \).

5.2 Choice and Negative Models

The node LIMP-FAULT in the model graph of the line (Fig. 2.4b) is an example of an implicit model: it serves only as a name for the disjunction of the two possible faults considered here, the normal short and the high-resistance short. The idea that sometimes it may be useful to work with a model that covers a number of other models motivates the following definition.

Definition 5.3 (Choice)
A set of models \( \{M_i\} \), is a choice for a model, \( M \), if

\[ M \Rightarrow M_1 \lor M_2 \lor \ldots \lor M_n. \]

As for the view (which can in fact be considered as "degenerate choice"), the contraposition indicates the effect of a choice in consistency-based diagnosis:

\[ \neg M_1 \land \neg M_2 \land \ldots \land \neg M_n \Rightarrow \neg M, \]

i.e. refuting all models of a choice for \( M \) suffices to refute \( M \).

Example 5.4
The most common example is the one mentioned to motivate the definition: a choice can be used to comprise the fault modes of one component in one model. For instance,

\[ L_{IMP-FAULT} \Rightarrow L_{IMP-SHORT} \lor L_{HI-RES-SHORT} \]

is a choice and represented by a hyper edge labelled "c" in Fig. 2.4b. We could specify a model for LIMP-FAULT by using the union of RL-IMP-

SHORT and RH-RES-SHORT:

\[ L_{IMP-FAULT} =
M(L, RL-IMP-SHORT \cup RH-RES-SHORT), \]

But note that this gives us (according to Lemma 4.2)

\[ M(L, RL-IMP-SHORT) \lor M(L, RH-RES-SHORT) \Rightarrow M(L, RL-IMP-SHORT \cup RH-RES-SHORT), \]
i.e. a view, and not a choice.

In general, the dual implication,

\[ M(C, R_1 \cup R_2) \Rightarrow M(C, R_1) \lor M(C, R_2), \]
is wrong. This is because a choice does not represent a result from set theory, but a statement about the physical reality, namely that the set \( \{M(C, R_i)\} \) really contains models for all (physically) possible behaviors covered by \( R_1 \cup R_2 \). Only if \( B_1, B_2, B_3 \) in Fig. 5.2 are really the only possible behaviors covered by the relation defining \( M \), the choice \( M \Rightarrow B_1 \lor B_2 \lor B_3 \) is correct. Of course, this may be an assumption rather than unconditionally true. In section 5.4 we will discuss this aspect. Choices may comprise only fault models of a certain type, and by introducing choices containing choices one can classify faults in a hierarchical manner.

![Diagram](image)

Figure 5.2 \( \{B_1, B_2, B_3\} \) is a choice for \( M \)

Example 5.5
There exists another choice in the model graph, namely

\[ L_{POSSIBLE} \Rightarrow L_{CORRECT} \lor L_{FAULTY} \]

LPOSSIBLE is another example of an implicit model which, by definition, covers all physically possible behavior modes of a line. But this implies that LPOSSIBLE is always true, and, hence, we have

\[ (\neg L_{FAULTY} \Rightarrow L_{CORRECT}) \land (\neg L_{CORRECT} \Rightarrow L_{FAULTY}) \]

Since this is a more general relation, we introduce the concept of a negative model in the following definitions.
Definition 5.4 (Negative Model)
M' is a negative model of M, if
\[ \neg M' \Rightarrow M. \]

Definition 5.5 (Negative Model Set)
\( \{ M_1, \ldots, M_n \} \) is a negative model set for M, if
\[ \neg M_1 \land \ldots \land \neg M_n \Rightarrow M. \]

This is equivalent to \( \{ M_i \} \) being a choice for a negative model for M and generalizes the "physical negation rule" used in GDE" ((Struss-Dressler 89)) for the exoneration of components: If the M; in Definition 5.5 are the models of all possible faults, then M must be a model of the correct mode, and the implication states "If the device is not broken, it is working".

The general concept of a negative model (set) appears to be rather weak. For relational models, removing only one value tuple \( v \in R \) that may occur in a real situation turns \( M(C, R) \) into a negative model \( M(C, R_{[v \in \emptyset]}) \) for itself. But we emphasize that the inferences in Definitions 5.4 and 5.5 are the only ones that allow our system to positively infer the validity of a model based on a detected inconsistency. In a view, \( M \Rightarrow M' \), the validity of \( M' \) is an implication of the validity of \( M \); but this only shifts the problem to the question of how to derive the validity of \( M \). Since the principle of consistency-based diagnosis is ruling out models based on contradictions, the link between a refuted model (or several of them) and a confirmed model as in Definitions 5.4 and 5.5 is central for a system that is to positively identify modes based on logical inferences.

In reality, besides the rejection of all possible faults, often sets of tests are used to identify a certain mode. Negative model sets can be used to introduce this element into consistency-based diagnosis.

5.3 Refinement

A model cannot only be a disjunction of other models as in a choice. Sometimes a conjunction of models establishes a new one. The example immediately associated with this idea is the replacement of a black-box model of an aggregate constituent by a set of models for its parts. But again, this is only one instance of a class of model relations which we call a refinement.

Definition 5.6 (Refinement)
A model set \( \{ M_1, \ldots, M_n \} \) is a refinement of a model, M, if
\[ M_1 \land M_2 \land \ldots \land M_n \Rightarrow M. \]

Example 5.6
If M represents the correct behavior mode of some aggregate component, and \( \{ M_i \} \) is the set of models of correct behavior modes of its subcomponents, then a refinement corresponds to a structural decomposition. In Example 4.13 we have already shown that the refinement relation holds for the relational models of the breaker and its parts, the distance protection, DP, and the switch, SW:
\[ M(DP, R_{DP-IMP}) \land M(SW, R_{SW}) \Rightarrow M(BR, R_{BR-IMP}). \]

A refinement states "If all parts work correctly then so does the entire system". Note that such an inference has some strong implications, such as exclusion of structural faults and of disturbances external to the model.

But did we not prove the refinement relation directly from the models of the breaker and its parts (in Example 4.13)? Yes, we did, but in order to do so, we implicitly identified two local variables to form one: the trip command, CMD. The principal problem is not that we ignored the line transmitting CMD. We could have introduced it together with two variables \( CMD_{DP} \) and \( CMD_{SW} \) which are equal if the transmission works. But still, the problem of setting local variables equal is only shifted to the interaction of the signal transmission line with the distance protection and with the switch, and the central and difficult problem remains: this step is based on the underlying assumption that the structure is unchanged and that, in particular, no unforeseen additional interaction occurs which may disturb communication between the ordinary constituents. In the DP philosophy such assumptions are typical candidates for diagnostic hypotheses, and we will return to this problem later in the discussion of simplifications.

Refinements work in both directions: having established that all \( M_i \) are true, one has shown that so is M. On the other hand,
\[ \neg M \Rightarrow \neg M_1 \lor \neg M_2 \lor \ldots \lor \neg M_n \]
holds, which means, for examples , that if the aggregate does not work, at least one of its components is broken.

To illustrate the effect in consistency-based diagnosis, consider a candidate, \( (C_1, C_2) \subset COMPS \). Imagine, \( C_1 \) and \( C_2 \) are aggregate components sharing one component, say \( C_{10} = C_{20} \). If the refinements are activated, i.e. inferences
\[ \lor C_{1i} \Rightarrow C_1 \land C_{2i} \Rightarrow C_2 \]
are added, then new conflicts are generated with the refinement sets replacing \( C_1 \) and \( C_2 \), respectively.
From the extended conflict set, the only single fault candidate generated under the focus of suspicion directed towards the refined level is \( C_{10} \). The point here is not so much, that this result can be obtained, but that this is achieved a) without any prediction based on models of the \( C_1 \) and \( C_2 \), and b) using the standard candidate generator without any changes.

**Example 5.7**

The utility of the refinement relation is not confined to structural decomposition. It can also express that we consider a set of physical aspects to characterize some behavior exhaustively. To illustrate this, consider again Example 5.3 where we split the correct model of the line into two distinct aspects. The relations describing these aspects, \( R_{L,IMP-Z} \) and \( R_{L,IMP-LV} \), link impedance and duration of the over current at the two ends of the line. Like in the case of the breaker's structural decomposition, we can embed the relations in the full representation \( (\mathcal{L}, \text{DOM}(\mathcal{L})) \) of the line. If we assume that an over current is present, i.e. we exclude the tuple \((0, 0, 0, 0)\) representing no or normal current, the relation \( R_{L,IMP} \) specifying the correct line is exactly the intersection of the two embedded relations,

\[
R_{L,IMP} = e_1(R_{L,IMP-Z}) \cap e_2(R_{L,IMP-LV}),
\]

and with Theorem 4.1 and Lemma 4.2 we have again established a refinement relation:

\[
M(L, R_{L,IMP-Z}) \land M(L, R_{L,IMP-LV}) \Rightarrow M(L, R_{L,IMP} \setminus \{(0, 0, 0, 0)\}).
\]

Note that, in this particular example, the implication holds also in the opposite direction, since the discussion in Example 5.3 showed that the models on the left-hand side are views of the entire model. So, we have

\[
M(L, R_{L,IMP-Z}) \land M(L, R_{L,IMP-LV}) \Leftrightarrow M(L, R_{L,IMP} \setminus \{(0, 0, 0, 0)\}).
\]

The potential value of a refinement for consistency-based diagnosis is twofold: First, its contraposition can create candidates on the refined level without instantiation of the respective models, as was illustrated above. Second, if there exists a way to derive the validity of all elements of a refinement (e.g. by means of fault models of the parts, or if the distinct physical aspects can be confirmed by tests), then the validity of the model on the right-hand side is obtained. However, the establishment of a refinement in both Example 5.6 and Example 5.7 is based on certain assumptions. This brings us to the concept of a simplification.

**5.4 Simplification**

In section 5.1 we introduced the concept of a view of a model. The validity of a view is a necessary condition for the validity of the model, and this is why we can use it as a substitute in consistency-based diagnosis. As an instance of a view we used

\( BR_{SIMPL+} \Rightarrow BR_{SIMPL} \).

It is often impossible to guarantee that the result of the simplification is still a model under all circumstances. In section 4, we pointed out that the danger of wrong conclusions when using a simplified model resides in the parts of the ideal behavior which are not covered by the simplified model, because if values in these parts are observed or inferred, this would lead to a (justified) refutation of the simplified model and an erroneous refutation of the respective behavior mode. Using the simplification nevertheless, is based on the assumption that in the situations we are looking at, we will not meet these dangerous areas of the relation. For instance, \( BR_{SIMPL+} \) is entailed by \( BR_{DIS} \) only in conjunction with the simplifying assumption about the level of the distance protection (namely restricting it to 1 or 2). We use DP's diagnostic hypotheses in order to formalize such a relation which we call model simplification:

**Definition 5.7 (Simplification)**

\( M' \) is a simplification of \( M \), iff

\[
\exists \{dhypt\} \subseteq \text{DHYT} \quad M \land \land \\text{dhypt} \Rightarrow M'.
\]

The effect of using a simplified model, \( M' \), is different from that of using a view. If \( M' \) is detected to be inconsistent with the observations, the same holds for \( M \) only under the condition that the underlying simplifying assumptions are true:

\[
\neg M' \land \land \text{dhypt} \Rightarrow \neg M.
\]

Stated differently, if \( M' \) is invalid, then so is \( M \), or (at least) one of the diagnostic assumptions is violated:

\[
\neg M' \Rightarrow \neg M \lor \lor \neg \text{dhypt}.
\]

We will analyze this aspect more formally in section 6.

In the examples discussed so far, we have encountered different types of simplifying assumptions. We will show that all of them can be adequately handled using the concept of a simplification.

**Example 5.8**

Simplification can be the result of a model transformation (as opposed to a representational transformation) as introduced in section 4.2. In example 4.13 we obtained the simplified model for the breaker, \( BR_{SIMPL+} \), from \( BR_{DIS} \) by "cutting off" the pieces of the specifying relation that correspond to the operation of the breaker for fault distances greater than 2 (normally the
cases when it acts as a back-up). The former model is a simplification of the latter in the sense of Definition 5.7:

$$BR_{DIST} \land \text{dhyPSIMPL} \Rightarrow BR_{SIMP} \land .$$

We can consider dhyPSIMPL to be just a name for the statement that no back-up situation is present:

$$\text{dhyPSIMPL} = \text{NO-BACK-UP}.$$ 

This is its domain-related meaning, its pragmatic aspect. But note that we are able to give it a precise meaning in terms of the relational models. Lemma 4.3 stated a sufficient condition for a justified use of the simplified model and provides us with a way to specify the simplifying assumption:

**Corollary 5.2**

Let $$R, R' \subseteq \text{DOM}(v_c)$$ and

$$\text{dhyP} \Rightarrow (\forall v_0 \in \text{DOM}(v_c)$$

$$(\exists s \in \text{SIT} \text{Val}(s, v_c, v_0) \Rightarrow v_0 \notin R \land R'))).$$

Then

$$M(C, R) \land \text{dhyP} \Rightarrow M(C, R').$$

This means, the diagnostic assumption suffices to exclude the occurrence of the particular values that would lead to a false refutation of the original model, $$M(C, R)$$. As a result of Example 4.13, we can define

$$\text{dhyPSIMPL} =$$

$$(\forall s \in \text{SIT} \neg \text{Val}(s, v_{BR\_DIST}, \text{OPEN}, 3, 2))$$

$$\land \neg (\text{Val}(s, \text{LEVEL}, \text{lev}) \land \text{lev} = 3)).$$

Note that, as opposed to the proposition NO-BACK-UP, this formula can be used and checked within the realm of the model-based prediction. For instance, we may detect that dhyPSIMPL contradicts values inferred by other models and, thus, gather evidence that this assumption may have to be retracted.

More formally and even stronger, Corollary 5.2 allows us to specify simplifying assumptions as they occur in model transformations in terms of a relation:

$$\text{dhyP} \Leftrightarrow$$

$$(\forall v_0 \in \text{DOM}(v_c)$$

$$(\exists s \in \text{SIT} \text{Val}(s, v_c, v_0) \Rightarrow v_0 \notin R \land R'))),$$

where

$$R \subseteq R' \subseteq R.$$ 

But this is the form of an impossibility, and, hence, of a relational model! We can write

$$\text{dhyP} = M(C, R),$$

or

$$\text{dhyP} = M(C, \text{DOM}(v_c) \setminus R').$$

Thus, we can state models and modeling assumptions in a very homogeneous way. To place this issue in the right perspective: often we may not want to expand the simplifying assumption into the form of a relational model. But we can treat it as an unspecified assumption first, and maintain the possibility to explicitly check it after it becomes suspect. An interesting aspect is that although the exact relation $$RR$$ may have about the complexity of the model we want to replace by a simplified version, also the simplifying assumption could be replaced by views with reduced complexity just as models can. More generally, the whole theory of model relations developed in this section can be applied to this type of modeling assumptions (In particular, we may simplify simplifying assumptions!)

**Example 5.9**

An important special model transformation is a projection to the nominal value or range of a certain variable, such as the normal range of the temperature of the environment, i.e.:

$$R' = R \cap [t_{low}, t_{high}] \times \text{DOM}(v_2) \times \ldots \times \text{DOM}(v_n).$$

In this case, dhyP represents the assumption that we encounter only situations with normal temperature conditions. In this case, we could explicitly state

$$\text{dhyP} = M(C, R'),$$

where

$$R' = [t_{low}, t_{high}] \times \text{DOM}(v_2) \times \ldots \times \text{DOM}(v_n).$$

In the discussion of refinements we also emphasized that they may be based on certain assumptions. Again, this can be expressed by an appropriate simplification:

$$\land M_i \Rightarrow M'$$

$$M' \land \text{dhyP} \Rightarrow M,$$

which means $$\{M_i\}$$ is considered as a refinement of a simplification of $$M$$, rather than a refinement of $$M$$. We may want to avoid the implicit model $$M'$$ and state directly

$$\land M_i \land \text{dhyP} \Rightarrow M.$$  

**Example 5.10**

Continuing on Example 5.6, we can now use the concept of a simplification to use the two partial models of the line (imposing constraints on impedance and duration of disturbance, respectively) conjunctively instead of the line model IMPEDANCE:

$$M(L, R_{L\_IMP, 2}) \land M(L, R_{L\_IMP, 1}) \land \text{dhyP} \Rightarrow M(L, R_{L\_IMP}),$$

where

$$\text{dhyP} = M(L, \text{DOM}(v_{L\_IMP})(0, 0, 0, 0)).$$

represents the assumption that an over current is present.

**Example 5.11**

The case of the structural refinement (Example 5.6) is different and, in fact, more complicated.
We pointed out that the relational models of the distance protection and the switch actually imply the relational model of the entire breaker system. Yet, we know that the correct functioning of the switch and of the protection system as single constituents is a necessary, but not a sufficient condition for the intended behavior of the aggregate; this is only guaranteed if the two constituents are combined in an appropriate way, and if there is no disturbing external influence. But why is this condition not vivid in the relational description? Why can't we make this assumption explicit in terms of relations as in the cases above? Since we can rigorously deduce

\[ M(D_{Pr}, R_{DP-Imp}) \land M(SW, R_{SW}) \Rightarrow M(BR, R_{BR-Imp}) \],

the structural soundness seems to be hardwired in the description. Indeed it is. It cannot be expressed in terms of excluded values of variables, because it is already effective in the choice of the variables, in this case, in using CMD as a variable in both the model of the switch and the protection system. If we turn the refinement into a simplification,

\[ M(D_{Pr}, R_{DP-Imp}) \land M(SW, R_{SW}) \land \text{dhp} \Rightarrow M(BR, R_{BR-Imp}), \]

the diagnostic assumption, \text{dhp}, is just a proposition stating that the structure is not violated. If we use different variables, \text{CMD}_{DP}, and \text{CMD}_{SW}, to describe the signal sent by the protection system and the signal received by the switch, then the structural soundness can (and has to be made explicit by the relation

\[ R_{CMD} = \{(\text{CMD}_{DP}, \text{CMD}_{SW}) | \text{CMD}_{DP} = \text{CMD}_{SW}\} \].

This allows us to specify the assumption that the intended contacts between constituents are working:

\[ \text{dhpcontact} = M(BR, R_{CMD}). \]

However, this solution forces us, on the one hand, to use the relation \text{R}_{CMD} explicitly when performing prediction on the subcomponent level, because otherwise the constituents remain isolated (The same effect would have been achieved by introducing the transmission line as a component with an equality constraint). On the other hand, this is still not an appropriate way to express that there are no additional contacts and influences, i.e., a kind of an explicit closed-world assumption in the diagnostic framework.

But still, we have a way to explicitly introduce the condition of structural soundness as a diagnostic assumption, thus providing a formal foundation for handling an intrinsic problem of model-based diagnosis.

So far, we were able to state simplifying assumptions in terms of relations in some appropriate representation. It is important to note that, in general, this is impossible for the simplifying assumptions underlying choices (including negative model sets). They are of a qualitatively different nature. If we state

\[ M \Rightarrow \exists M_i, \]

we normally exclude a large number of theoretically possible models. Consequently, if we would like to specify a simplifying assumption supporting a choice:

\[ M \land \text{dhp} \Rightarrow M', \]

\[ M' \Rightarrow \exists M_i, \]

or briefly,

\[ M \land \text{dhp} \Rightarrow \exists M_i, \]

then \text{dhp} stands for ruling out any model not stated, which is a big number of relations. The underlying reason is that this model relation, in contrast to others, is not based on the assumption that certain (sets of) values in \text{DOM}(\text{y}_{_M}) are not encountered. Rather, it is motivated by the physics of the system to assume that certain relations cannot be present. So, rather than excluding one relation, one has to exclude a (huge) set of relations (not equivalent to excluding the union of these relations!). This is not feasible in general. Hence, in the case of a choice, simplifying assumptions will be merely stating that no other behavior models are possible, or not considered at the present time.

Example 5.12

If we want to use a negative model set carefully, we can turn the completeness axiom into a simplification assumption, e.g., by stating

\[ \neg M' \Rightarrow M \]

\[ M' \land \text{dhp} \Rightarrow M_1 \lor M_2 \lor \ldots \lor M_n, \]

i.e. making \{M_i\} a choice for a simplification of a negative model for M. This allows us to introduce a retractable completeness assumption for fault models and tests. For instance, the fault model of the line (Example 5.4) could be stated as a simplification

\[ L_{\text{IMP-FAULT}} \land \text{dhypfmc} \]

\[ \Rightarrow L_{\text{IMP-SHORT}} \lor L_{\text{HL-RES-SHORT}}, \]

where \text{dhypfmc} ("Fault models complete") represents the assumption that the normal short and the high-resistance fault are all ways of a line to fail. From this, we derive

\[ \neg L_{\text{IMP-SHORT}} \land \neg L_{\text{HL-RES-SHORT}} \]

\[ \land \text{dhypfmc} \Rightarrow \neg L_{\text{IMP-FAULT}}. \]
This suggests that the diagnostic assumption can also be interpreted as the negation of unknown faults, and, thus, we have a way to handle unknown faults in the DP framework: the absence of unknown faults can be introduced as a working hypothesis which might be abandoned.

**Remark 5.3**

Clearly, Definition 5.7 captures the transitivity of simplification: if $M'$ is a simplification of $M$, and $M''$ is a simplification of $M'$, then $M''$ is a simplification of $M$.

Moreover, the simplification property carries over to views:

**Remark 5.4**

If $M'$ is a simplification of $M$, and $M''$ is a view of $M'$, then $M''$ is a simplification of $M$:

$(M \land \text{dhyp} \Rightarrow M' \land M' \Rightarrow M'') 
\Rightarrow M \land \text{dhyp} \Rightarrow M''$.

5.5 The Model Graph

Through the definitions in the preceding subsections, we have introduced a set of relations between models and discussed their role in consistency-based diagnosis. In particular, we demonstrated that operations on models quite different in terms of their programmatic aspects and physical interpretation can lead to the same logical relations. This is important for precisely understanding the impact of the pragmatically motivated structuring of the models. For instance, we have shown the common logical kernel of structural and behavioral refinement. And we demonstrated that a simplification can represent quite different modeling assumptions, such as the completeness of the faults modelled, the unviolated structure of an aggregate, and appropriateness of certain modifications or approximations of the descriptive relation.

In Fig. 5.3a, b the model relations of the breaker and the line discussed in the examples are summarized, except for the refinements. Compared to the informally introduced links in the graph of Fig. 2.4b, we have reproduced the structure and given the links precise meanings. For the sake of clarity, we have added the ideal behavior modes as implicit models (in bold characters). For practical purposes, this is not necessary. One can identify the ideal mode with the best view (i.e. the one that is not a view of some other view of this model), unless one cautiously wants to regard even the best model only as a simplification. o-arcs correspond to simplifications, v-arcs are views, and choices are marked with "c".

The graph indicates, for instance, that the models BRSIMPL+ and BRSIMPL are valid models for the correct behavior of the breaker under the simplifying assumption NO-BACK-UP. And we have to consider LIMP-SHORT as the only possible fault, given the assumption NO-HI-RES.

\[
\begin{align*}
\text{BRCORRECT} & \xrightarrow{v_1} \\
\text{BRIMPEDANCE} & \xrightarrow{v_{DIST}} \\
\text{BRDIST} & \xrightarrow{v_{NO-BACK-UP}} \\
\text{BRSIMPL+} & \xrightarrow{v_{LEVEL}} \\
\text{BRSIMPL} & \\
\text{LPOSSIBLE} & \\
\text{LCORRECT} & \xrightarrow{v_2} \text{LFAULTY} \xrightarrow{v_3} \\
\text{LIMPEDANCE} & \xrightarrow{v_{DIST}} \text{LIMP-FAULT} \\
\text{LDIST} & \\
\text{LIMP-SHORT} & \xrightarrow{v_{DIST}} \text{LHI-RES-SHORT} \\
\text{LSHORT} & \\
\end{align*}
\]

Figure 5.3a Part of the model graph of the model graph of the breaker

Figure 5.3b Part of the model graph of the line

The model graph is used to guide the selection and instantiation of models (and also their deactivation) in the course of the diagnostic process. Basically, we start at the leaves of the model graphs of components; views and simplifications are to be used...
first in order to save costs in prediction. We climb up in the model graph if there is evidence that a revision of modeling assumptions and/or a more detailed model is required.

As a technical remark, we state that the ideal behavioral modes define the places where assumptions are introduced (namely that the respective mode is the actual one) which are then recorded by the ATMS and propagated via view and simplification links. It is only simplification links that add further assumptions (which may be either unspecified ATMS assumptions or nodes ultimately justified by explicit simplification conditions as specified by Corollary 5.2). For instance, the model LDIFF of line 11 is labelled \{CORRECT(l1)\}, and inferences based on the model BRSIMP+ of br1 are based on the assumption set
\[ \{\text{CORRECT(br1)}, \text{NO-BACK-UP(br1)}\} \].
Thus, if a prediction of BRSIMP+ contradicts predictions based on some assumption set A, then
\[ A \cup \{\text{CORRECT(br1)}, \text{NO-BACK-UP(br1)}\} \]
will be recorded as a conflict.

As a side-effect, this allows us to use ATMS-based focusing techniques as described in [Dressier-Farquhar 90] for letting simplification assumptions guide the focus for prediction.

6 Diagnosis with Multiple Models

In this section, we present a strategy that uses the model graph in the intended way. It is meant to be one example out of a family of possible strategies and reflects some special characteristics of the network domain (for instance structural refinement is not used in DPNet, and observations are either free as part of the messages or impossible). We will first provide some foundations concerning the structure of the system description and its selective use, then describe the strategy, and finally we apply it to a set of cases in the power transmission network.

6.1 The Foundation

With the introduction of the model graph, the system description, SD, is no longer fixed for the entire diagnostic process, but may change. SD can (and for practical purposes has to) be decomposed into different knowledge sources. The main distinction is the one between knowledge about
- the application domain and
- the specific device to be diagnosed.

The device description is, according to the philosophy of the approach, only a structure description, D-STRUCTURE (although, in practice, one might want to, or have to, incorporate additional models, e.g. describing the particular teleology of aggregates).

The domain knowledge comprises knowledge about
- the representations used to define models, including
  - the variable domain, i.e. the set of values,
  - discrepancy criteria ("What is considered to be contradictory?")
  - transformations between different representations,
- the constituents in the application domain (LIBRARY), containing
  - the model graphs for the constituents (M-GROUP), i.e. relations as introduced in section 5, and
  - the set of model definitions for the explicit models (M-DEF).

The term model definition refers to the "predictive body" of an explicit model (as opposed to its name occurring in the model graph) such as a set of constraints or rules. In the relational case, it has to implement a model according to Definition 4.1. Table 6.1 summarizes this structuring of SD and illustrates the contents of the various elements logically by examples.

<table>
<thead>
<tr>
<th>APPLICATION DOMAIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>REPRESENTATION</td>
</tr>
<tr>
<td>VARIABLE DOMAINS</td>
</tr>
<tr>
<td>e.g. ( y_0 \in \text{DOM}(y_c) )</td>
</tr>
<tr>
<td>( \Rightarrow y_0 = y_1 \lor y_0 = y_2 \lor y_0 = y_3 \lor \ldots )</td>
</tr>
<tr>
<td>DISCREPANCIES</td>
</tr>
<tr>
<td>e.g. ( \forall s \in \text{SIT} )</td>
</tr>
<tr>
<td>( \text{Val}(s, y_c, y_o) \Rightarrow \neg \text{Val}(s, y_c, y'_o) )</td>
</tr>
<tr>
<td>TRANSFORMATIONS</td>
</tr>
<tr>
<td>e.g. ( \forall s \in \text{SIT} )</td>
</tr>
<tr>
<td>( \text{Val}(s, y_c, y_o) \Rightarrow \text{Val}(s, y'_c, y'_o) )</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>BEHAVIOR MODELS</td>
</tr>
<tr>
<td>MODEL GRAPHS</td>
</tr>
<tr>
<td>e.g. ( M \land \text{dhyp} \Rightarrow M' )</td>
</tr>
<tr>
<td>MODEL DEFINITIONS</td>
</tr>
<tr>
<td>e.g. ( M \Leftrightarrow \forall y_0 \in \text{DOM}(y_c) )</td>
</tr>
<tr>
<td>( (\exists s \in \text{SIT} \text{ Val}(s, y_c, y_o) \Rightarrow y_0 \in \text{R} ) )</td>
</tr>
<tr>
<td>DEVICE</td>
</tr>
<tr>
<td>STRUCTURE</td>
</tr>
<tr>
<td>e.g. ( \forall s \in \text{SIT} )</td>
</tr>
<tr>
<td>( \text{Val}(s, y_c, y_o) \Rightarrow \text{Val}(s, y_d, y_o) )</td>
</tr>
</tbody>
</table>

Table 6.1 Elements of SD

The strategies we have in mind aim at exploiting only a part of the possible predictions. At each stage of the diagnostic process, SD is given by
\[ \text{SD} = \text{REPRESENTATION} \cup \text{D-STRUCTURE} \cup \text{M-GROUP} \cup \text{M-DEF} \]
\[ = \text{SD\_CORE} \cup \text{M-DEF\_ACT}, \]
where the set of active models, M-DEF \_ M-DEF, normally contains only a small subset of model definitions at a time. For the sake of simplicity, we assume that SD\_CORE is always complete, and remains stable (actually, we need to consider only subsets of it, as well, dependent on the active models). Although M-DEF \_ M-DEF will also reflect a structural focus (for instance, DPNet gains its efficiency from analyzing only a small environment of the short circuit within the overall network), this is not the topic of this paper (see [de Kleer 91], [Dressler-Farquhar 90], [Freitag-Friedrich 91]).

Here, we are interested in selecting an appropriate layer in the model graphs of the interesting constituent elements guided by the principle "As simple as possible, as detailed as necessary".

Organizing the use of models in the diagnostic process according to this principle is based on monotonicity properties of changes in the space of diagnoses whose ultimate foundation is captured by the following theorems. They tell how the space of diagnoses is changed as a result of model switches along the various edges in the model graph.

**Theorem 6.1**

Let \( \{M_i \} \) be a refinement of \( M \), i.e.
\[
\bigwedge M_i \Rightarrow M^f,
\]
and \( M^f \) and \( \{M\_DEF(M_i)\} \) be the respective model definitions.

If \( \delta \) is a diagnosis for
\[
OBS \cup SD\_CORE \cup M\_DEF\_ACT \cup \{M\_DEF(M_i)\}
\]
then \( \delta \) is a diagnosis for
\[
OBS \cup SD\_CORE \cup M\_DEF\_ACT \cup \{M\_DEF(M^f)\}.
\]

Note that a view,
\[
M \Rightarrow M^f
\]
can be regarded as a "degenerate refinement" with one element. Thus, we obtain the following corollary.

**Corollary 6.2**

Let \( M^f \) be a view of \( M \), i.e.
\[
M \Rightarrow M^f,
\]
and \( M^f \) and \( M\_DEF(M) \) be the respective model definitions.

If \( \delta \) is a diagnosis for
\[
OBS \cup SD\_CORE \cup M\_DEF\_ACT \cup \{M\_DEF(M)\}
\]
then \( \delta \) is a diagnosis for
\[
OBS \cup SD\_CORE \cup M\_DEF\_ACT \cup \{M\_DEF(M^f)\}.
\]

Basically, this says that when working with a view, \( M^f \), of a model, \( M \), we do not miss a diagnosis we would obtain when using the original model, \( M \). Stated differently, switching from the view to the more powerful model \( M \) is a step that produces a subset of the diagnoses derivable from \( M\_DEF(M^f) \). Intuitively, this is clear, because \( M\_DEF(M) \) entails more (or more fine-grained) predictions, reveals more discrepancies, and hence, rules out more diagnoses. The theorem provides the ultimate justification for applying models, which are gained by (qualitative) abstraction or which model only particular physical aspects, in order to cut down the space of possible diagnoses prior to further investigation with more fine-grained, but also more costly models.

Of course, when using simplified models, this kind of monotonicity will be restricted, as indicated by the following theorem.

**Theorem 6.3**

Let \( M^f \) be a simplification of a refinement \( \{M_i\} \) of \( M \):
\[
\bigwedge M_i \Rightarrow M \land dhyp \Rightarrow M^f,
\]
and \( \{M\_DEF(M_i)\} \) and \( M\_DEF(M^f) \) be the respective model definitions.

If \( \delta \) is a diagnosis for
\[
OBS \cup SD\_CORE \cup M\_DEF\_ACT \cup \{M\_DEF(M_i)\}
\]
then \( \delta \) is a diagnosis for
\[
OBS \cup SD\_CORE \cup M\_DEF\_ACT \cup \{M\_DEF(M^f)\},
\]
or \( dhyp \delta \).

This theorem formulates what is in accordance with our intuition, namely that by using a simplified model, the system will infer all diagnoses that can be obtained from the original one and do not contain the retraction of the underlying simplifying assumption.

Part of the diagnosis space that is based on \( \neg dhyp \) may be invisible; however, it can be regained in DP, since the diagnostic assumptions are made explicit and kept in dependencies. Note that, as soon as the set of diagnoses obtained with the simplified model shrinks to 0 or is considered implausible due to some general or domain specific criterion, this forces the retraction of diagnostic hypotheses.

**Proof of Theorems 6.1 and 6.3**

We can regard a refinement and Theorem 6.1 as a special case of a simplification and Theorem 6.3, respectively, with \( dhyp=True \) (i.e. \( M \) is an unconditional implication of \( \bigwedge M_i \)). In this case, no diagnosis can contain \( dhyp \). Hence, we only need to prove Theorem 6.3.

Assume \( \delta=\delta_{COMPS} \cup \delta_{DHYP} \) is a diagnosis for
\[
OBS \cup SD\_CORE \cup M\_DEF\_ACT \cup \{M\_DEF(M_i)\},
\]
but not for
\[
OBS \cup SD\_CORE \cup M\_DEF\_ACT \cup \{M\_DEF(M^f)\}.
\]

The latter statement means that
\[
OBS \cup SD\_CORE \cup M\_DEF\_ACT \cup \{M\_DEF(M^f)\} \cup FAULTY(C) \cup CORRECT(C) \cup \neg dhyp \cup dhyp \delta_{DHYP} \delta_{DHYP}
\]
is inconsistent. If dhyp were not in δ, then 
\[ \text{dhyp} \notin \text{DHYP}, \] 
and, because SD\text{CORE} contains \( \wedge M_i \wedge \text{dhyp} \Rightarrow M' \), 
\[ \text{OBS} \cup \text{SD\text{CORE}} \cup M-\text{DEF}_\text{ACT} \cup \{M-\text{DEF} (M_i)\} \] 
\[ \cup \cup \text{FAULTY}(C) \cup \text{CORRECT}(C) \] 
\[ C \in \text{COMP}\text{S} \cup \text{COMP}\text{S} \) \text{COMP}\text{S} \cup \text{DHYP} \] 
\[ \text{dhyp} \notin \text{DHYP} \cup \text{dhyp} \] 
would be inconsistent, which contradicts the assumption that \( \delta \) is a diagnosis when \( \{M-\text{DEF}(M_i)\} \) is used.

What is the basis for utilizing negative models (including fault models)? As opposed to views, refinements, and its simplifications, they are not meant to replace but to complement the model they are negative models for. Hence, the answer to the question is the trivial statement that the set of diagnoses (according to Definition 3.1) shrinks monotonically when M-\text{DEF}_\text{ACT} grows monotonically, because more predictions reveal more inconsistencies and, hence, rule out more mode combinations.

**Lemma 6.4**

If \( \delta \) is a diagnosis for 
\[ \text{OBS} \cup \text{SD\text{CORE}} \cup M-\text{DEF}_\text{ACT} \cup \{M-\text{DEF}(M)\} \] 
then \( \delta \) is a diagnosis for 
\[ \text{OBS} \cup \text{SD\text{CORE}} \cup M-\text{DEF}_\text{ACT}. \]

In particular, \( M \) may be a fault model for a model in M-\text{DEF}_\text{ACT}, and it may eliminate some of the diagnoses obtained without activating \( M \), specifically some diagnoses containing the constituent \( M \) is associated with. In [Struss-Dressler 89] a simple example demonstrates that the effect can be significant.

With choices, 
\[ M \Rightarrow V M_i \] 
there may be two cases:

Either \( M \) is only an implicit model, just introduced to represent the choice, and M-\text{DEF}(M) is empty. Then expanding it, i.e. adding \( \{M-\text{DEF}(M_i)\} \) to M-\text{DEF}_\text{ACT} reduces the space of diagnoses according to Lemma 6.4.

If, on the other hand, we want to attach a model definition to \( M \), then it is only reasonable, if it establishes \( M \) as a view for all \( M_i \): 
\[ \forall i M_i \Rightarrow M. \]
This is covered by Corollary 6.2 which then states that the expansion of the choice will reduce the space of diagnoses.

With the theorems and lemmata stated above, we have achieved our goal of characterizing the changes in the space of diagnoses as a result of switching models along the different links.

**6.2 The Strategy**

In order to define a strategy for navigating through the model graph, we assume there exists a module that filters the theoretically possible diagnoses according to some criteria (likelihood, cardinality, etc.) in order to generate the set \( \Delta \) of candidates actually under consideration. This filter is the focus of suspicion introduced in section 3.

The strategy loops through the following steps:

1. **(Re-)Compute \( \Delta \)**

2. If \( \Delta = \emptyset \) (i.e. no accepted diagnosis), then retract assumptions, i.e.
   if there are any simplifying assumptions to retract
   then include (a subset of) them in the focus of suspicion and recompute \( \Delta \).
   If \( \Delta = \emptyset \), then deactivate the models based on the simplifying assumptions occurring in \( \Delta \) and activate the unsimplified ones; go to 1 .
   else go to 2 .
   else return "diagnosis impossible" .

3. If \( |\Delta| = 1 \) (i.e. a single leading candidate)
   then check leading candidate: i.e.
   if there are view links to move up
   then deactivate views and activate their stronger models
   else if there are inactive fault models (under the current simplifications) of components occurring in \( \Delta \), then activate the fault models; go to 1 .
   else return \( \Delta \).

4. If \( |\Delta| > 1 \) (i.e. competing candidates)
   then discriminate, i.e.
   if there are further potentially informative observations,
   then select and enter observations; go to 1 .
   else check each candidate as if it were a single candidate (step 3).

In step 4, an instance of the generic measurement proposer (IFreitag 90) is used. In general, it is desirable to weaken the priority of observations over discrimination by exploiting more models; however, in DPNet observations are cheap as part of the already available message burst.

In step 2, as an alternative to considering the diagnosis having failed, one could try to weaken the
filter on candidates, for instance, by increasing the maximal cardinality of admissible diagnoses.

For step 2 and 3, we assume criteria for selecting simplifying assumptions and views. But note that the diagnostic process itself provides criteria: because of dependency recording of diagnostic hypotheses, the system checks whether a retraction can fill \( \Delta \) again, without actually having to activate the respective models.

### 6.3 An Example

As an illustration of the strategy, we consider the following examples from the transmission network domain (Fig. 6.1):

- **Case 1**: There is a short on \( b_1 \); breakers \( br_1 \) and \( br_4 \) open correctly at level 2, while the other breakers remain closed.
- **Case 2**: There is short on \( l_2 \); \( br_4 \) opens correctly at level 1; but \( br_3 \) does not, and \( br_1 \) intervenes as a backup on level 3.
- **Case 3**: There is a high-resistance short on \( l_2 \), again \( br_3 \) fails, and \( br_1 \) and \( br_4 \) open at level 3 and 2, respectively.

Note that all three cases produce the same pattern of open breakers and differ only in terms of intervention levels.

In the following description of the application of the strategy, the focus of suspicion allows only minimal diagnoses with at least one short circuit and at most one faulty breaker. We treat this cardinality filter as fixed, although, in DPNet, it can be weakened dynamically.

Let us look at case 1. We start with a focus of suspicion that does not include any simplifying assumptions. Based on the messages of the tripped breakers and the leaf models of the correct modes, i.e. \( BR_{SIMPL} \), \( L_{DIST} \), and the respective model for the bus-bars, two minimal conflicts are detected (For convenience, we abbreviate \( CORRECT(C) = C \):

\[
\begin{align*}
CO_1 &= \{ br_1, NO-BACK-UP(b_{r1}), l_1, b_1 \}, \\
CO_2 &= \{ br_4, NO-BACK-UP(b_{r4}), l_2, b_1 \}.
\end{align*}
\]

Because the focus of suspicion excludes candidates with breakers only (such as \( \{br_1, br_4\} \)), and candidates with diagnostic assumptions (e.g. \( \{NO-BACK-UP(b_{r1}), l_2\} \)), the following minimal diagnoses are generated:

\[
\{b_1\}, \{l_1, br_4\}, \{l_2, br_1\}, \{l_1, l_2\}.
\]

Because no further observations help, \( BR_{SIMPL} \), which additionally makes predictions about the level, is activated for discrimination. All four minimal candidates survive this step. Next, fault models \( L_{SHORT} \) and \( B_{SHORT} \) are activated. Because

---

Figure 6.1 The fault examples. Tripped breakers are annotated with the respective level, missing interventions indicated by "!!"
L_{SHORT} is based on the simplifying assumption NO-HI-RES, the following conflicts are obtained:

\{ \text{SHORT}(l_1), \text{NO-HI-RES}(l_1), \text{br}_2, \text{NO-BACK-UP}(br_2) \},
\{ \text{SHORT}(l_2), \text{NO-HI-RES}(l_2), \text{br}_3, \text{NO-BACK-UP}(br_3) \}.

This corresponds to the detection of one failing breaker for each faulted line, since \text{br}_3 and \text{br}_2 should protect \text{l}_2 and \text{l}_1, respectively. From these conflicts,

\text{br}_2 \land \text{NO-BACK-UP}(br_2) \land \text{NO-HI-RES}(l_1) \Rightarrow l_1,
\text{br}_3 \land \text{NO-BACK-UP}(br_3) \land \text{NO-HI-RES}(l_2) \Rightarrow l_2,

and two new positive conflicts are derived:

\text{CO}_3 = \{ \text{br}_1, \text{NO-BACK-UP}(br_1), \text{br}_2, \text{NO-BACK-UP}(br_2), \text{NO-HI-RES}(l_1), b_1 \},
\text{CO}_4 = \{ \text{br}_3, \text{NO-BACK-UP}(br_3), \text{NO-HI-RES}(l_2), b_1 \}.

The only remaining candidate is

\{ b_1 \},

because the other three candidates have to be extended, and are filtered out by the double fault focus. The fault on bus-bar \text{b}_1 is correctly determined.

In case 2, we obtain the same initial candidates

\{ b_1, l_1, l_2 \}, \{ l_1, b_1 \}, \{ l_1, l_2 \}, \{ l_2, br_4, b_1 \}, \{ l_1, l_2 \}.

In the second step, however, \text{NO-BACK-UP}(br_1) directly contradicts level 3 at \text{br}_1, and, through information about the level at \text{br}_4, a conflict involving \text{br}_4 and \text{l}_2 is detected. The first conflict,

\text{CO}_1 = \{ \text{NO-BACK-UP}(br_1) \},

replaces \text{CO}_1 (because it is a subset), and the second one,

\text{CO}_2 = \{ \text{br}_4, \text{NO-BACK-UP}(br_4), l_2 \},

subsumes \text{CO}_2. No admissible candidates can be generated; \Delta is empty, and the system has to retract assumptions. The conflicts obtained so far indicate which simplifying assumptions are candidates for retraction (obviously, \text{NO-BACK-UP}(br_1)). For instance, \text{NO-HI-RES}(l_1) does not appear in a minimal conflict and is not suggested for retraction. The focus of suspicion is extended in a way such that one simplifying assumption out of the minimal conflicts may occur in a candidate while the other restrictions are maintained. New candidate generation from the existing conflicts \text{CO}_1, \text{CO}_2 (i.e. without any prediction) under this new focus produces only the candidate

\{ \text{NO-BACK-UP}(br_1), l_2 \}.

This indicates that there is no use in retracting, for instance, \text{NO-BACK-UP}(br_4), because this would not lead to acceptable candidates. \text{NO-BACK-UP}(br_1) is retracted and \text{BR}_{BSTR} is activated, the model that considers fault distances greater than 2. The step reveals a new conflict involving \text{br}_1, \text{l}_2, and other lines and bus-bars, thus confirming the previous diagnosis. In the next step (check), as in case 1, the instantiated fault model of \text{l}_2 contradicts the correct model of breaker \text{br}_3, and from

\text{br}_3 \land \text{NO-BACK-UP}(br_3) \land \text{NO-HI-RES}(l_2) \Rightarrow l_2,

and the respective conflicts the candidate

\{ l_2, \text{BR}_{BSTR}, \text{NO-BACK-UP}(br_1) \},

is obtained, which is correct and also remains valid after the check with the impedance models.

In case 3 (the high resistance short on \text{l}_2), the strategy also generates

\{ l_2, \text{br}_3, \text{NO-BACK-UP}(br_1) \},

as before, but the final check with the impedance fault model of the line reveals a discrepancy: levels 2 and 3 for \text{br}_4 and \text{br}_1, respectively, imply impedances corresponding to 1.5 and .85 line lengths minimum which is inconsistent with any of the fault models under the assumption NO-HI-RES. Retracting it forces the system to go back to \text{L}_{IMP-FAULT} and consider its full choice, now including \text{L}_{HI-RES}-\text{SHORT}. As the only possible diagnosis,

\{ l_2, \text{NO-HI-RES}(l_2), \text{br}_3, \text{NO-BACK-UP}(br_1) \}

remains. In other words, there is a high-resistance short on \text{l}_2, \text{br}_3 failed and \text{br}_1 acted as a back-up, which is the correct diagnosis.

Thus, we illustrated that the system solves the normal case quite efficiently and that it can fall back to more powerful and complete models, if necessary. Furthermore, we have demonstrated that information contained in the conflicts provides a basis for focused retraction of diagnostic assumptions, and that we are not solely dependent on predefined preferences for the assumptions.

7 Discussion

The theory of multiple models we developed, together with the capabilities of the DP framework enables us to structure our knowledge about a system's behavior in such a way that we can obtain results by instantiating and using only a portion of the entire model. Furthermore, since modeling assumptions can be represented explicitly, the system can reason about them and has a basis for a controlled navigation through the model graph. The example demonstrates the twofold goal of the system:

- **Efficiency** is gained for the simple standard cases. This is achieved by using simplifying modeling assumptions which reflect the properties of these cases.

- **Completeness** of the model-based approach has not been sacrificed, because the application of simplified models is performed as defeasible inferences. Information provided by the diagnostic
process (represented in terms of conflicts and candidates) can be exploited in order to select the assumptions appropriate for retraction.

In order to relate this approach to other work in the field that aims at more efficient diagnostic procedures, imagine that we have to diagnose a system with n components, each with m possible faults, and the presence of k simultaneous faults has to be considered. In an exhaustive search of this space of complete model combinations, prediction (i.e. the application of models) would cost in the order of

\[ n^c \cdot (m^*n)^k, \]

where c is the cost of executing a model of a single component, and, hence, \( n^c \) the cost of one system model. Current heuristics and control techniques basically aim at diminishing

- k (for instance, based on information about the cardinality of the most plausible candidates)
- n in the term \( m^n \) (for instance, by considering only fault models of components occurring in leading diagnoses)
- n in the term \( n^c \) (for instance, by focusing on partial system structures that have been detected to contain a faulty component),
- n in \( n^c \) by implementation techniques for caching results of model execution (e.g. as a side-effect of employing an ATMS).

Additionally,

- \( (m^*n)^k \) can be reduced (for instance, by terminating the search after a sufficient set of diagnoses has been found).

What these techniques for focusing in structure leave untouched, is c, the cost of using a single model. Introducing multiple models, differing in the level of abstraction and the simplifying assumptions involved, aims at providing the system with models which are cheaper to use.

Also m, the number of fault models, is affected: simplification can mean ignoring certain kinds of faults, and abstraction can collapse a number of fault models into a single one. As a side-effect of model abstraction, caching becomes more effective: if distinctions between certain values of variables are eliminated, and, hence, the cardinality of the domain is reduced, it is more likely that a model is applied to the same vector of values again. Note also, that a more detailed model may not only increase the cost of inferences but also potentially require more variables to be measured which may be impossible or expensive to obtain.

One might argue that the overhead required for controlling model switching may overwhelm the efficiency gained from the use of simpler models. The important aspect is that we obtain diagnoses for the majority of simple standard cases (like case 1 in section 6.3) very efficiently, while spending more time on solving the difficult and rare problems is justifiable. In other words, simplifications should only be introduced if they really apply to a significant number of cases. Another consideration concerns the fact (or expectation) that even the difficult cases will become extraordinary only with respect to a subset of the existing assumptions, thus allowing the system to exploit many of them.

But then the development of powerful control schemes for determining the appropriate set of violated assumptions in a focused and effective manner becomes crucial. As stated above, supporting information can be extracted from the diagnostic procedure itself. This distinguishes our approach from that of [Falkenhainer-Forbus 92], which requires predefined statements about what to consider, given certain queries. [Dressler-Boetcker 92] tackles the problem as an instance of belief revision.

Automated model generation through abstraction and simplification becomes another important task (see e.g. [Weld 90], [Weld-Addanki 92]). We believe that the relational representation presented in section 4 forms a natural and convenient basis for this. It requires only the analysis of sets and mappings between them (as opposed to a theorem prover, for instance). The task gains importance not only because we have to build and maintain more models for the framework presented. There will often be a number of orthogonal dimensions for simplification and abstraction which carry the danger of combinatorial explosion of the model graph and/or set the problem of deciding which combinations of simplifications are really relevant. Furthermore, some of the possible simplifications may be due to characteristics of a particular device rather than an intrinsic feature of the single component model. This threatens the fundamental idea of having a complete model library for a whole domain. At least, it requires possibilities for automating model generation.

Another element that has to be developed is a theory of costs of models as a basis for model selection. In addition to model relations presented here, it has also to integrate other model switching actions, such as structural refinement and finally relate to considerations about costs and gain of measurements, tests, and repair actions.
Despite these open issues, we believe that the extension of consistency-based diagnosis to multiple models is an important step towards a more general diagnostic theory as well as to systems mature enough for significant and complex applications as is demonstrated by the non-trivial domain of DPNet.

Acknowledgements
I would like to thank Toni Beschta, Danny Bobrow, Stefano Cermignani, Mike Chantler, Johan de Kleer, Oskar Dressler, Hartmut Freitag, Gerhard Friedrich, Georg Gottlob, Roy Leitch, Michael Montag, Wolfgang Nejdl, Monika Pfau, Olivier Raiman, Alberto Stefanini, Giorgio Tornielli, Brian Williams, researchers in the BEHAVIOR project, and several reviewers for discussions and comments on earlier versions. I am indebted most to Walter Hamscher who struggled with various drafts and versions at different occasions and provided important help. This work was supported in part by BMFT (ITW 8506 E4, ITW 9001 A9) and CEC (P5143, 3249).

References

[Beschta et al. 92]

[Davis 82]
Davis, R., Expert Systems: Where Are We? And Where Do We Go From Here?. The AI Magazine, Spring 1982

[Davis 84]

[de Kleer-Mackworth-Reiter 90]

[de Kleer 91]

[de Kleer-Williams 87]

[de Kleer-Williams 89]
de Kleer, J., and Williams, B. C., Diagnosis with Behavioral Modes. In: Proc. 11th Int. Joint Conf. on Artificial Intelligence, Detroit, MI, 1989

[Dressler-Boettcher 92]

[Dressler-Farquhar 90]

[Dressler-Struss 92]

[Falkenhainer-Forbus 92]

[Freitag-Friedrich 91]

[Friedrich-Gottlob-Nejdl 90]

[Hamscher 90]

[Holtzblatt 88]

[Preist-Welham 90]
[Raiman 89]
Raiman, O., *Diagnosis as a Trial*. IBM Scientific Center, Paris 1989

[Reiter 87]

[Struss 88a]

[Struss 88b]

[Struss 88b]

[Struss 89a]
Struss, P., *Diagnosis as a Process*. First International Workshop on Model-Based Diagnosis, Paris, 1989

[Struss 89b]

[Struss 91]

[Struss-Dressler 89]
Struss, P., Dressler, O., *"Physical Negation" - Integrating Fault Models into the General Diagnostic Engine*. In: Proc. 11th Int. Joint Conf. on Artificial Intelligence, Detroit, MI, 1989

[Weld 90]

[Weld-Addanki 92]