1 Introduction

Diagnosis in general requires more than just passively observing the behavior of a faulty system. Often the only way to gain useful information is to perform a test series: Physicians do not only listen to a description of symptoms but examine their patients; technicians actively locate the faulty component of a malfunctioning device. The observations made in the course of such experiments form the basis for a successful diagnosis.

Active diagnosis therefore requires to reason about more than a single state of the system to be examined. We propose to capture this aspect by appealing to Action Theory. A system is specified by its components and the way these entities change their states when being manipulated by means of actions. Performing diagnosis then amounts to appropriately interpreting observations concerning a system's state prior, during, and after the execution of a series of actions. Additionally, using Action Theory as the formal basis for dynamic diagnosis may help with finding further actions to be taken towards fully determining the cause of an observed system failure.

As an example for dynamic diagnosis consider the electric circuit depicted in Figure 1. It involves a number of components, some of which—several switches—can be directly manipulated by actions. Other components might be indirectly affected. It is assumed that only some components are directly observable. Now, suppose we close switch $\mathtt{s}_1$ in the current state depicted and we observe that afterwards light bulb $\mathtt{li}$ is still off. This calls for diagnosis: Under normal circumstances, relay $\mathtt{re}_1$ should have become activated and attracted switch $\mathtt{s}_2$, which in turn should have activated relay $\mathtt{re}_2$ and, hence, closed switch $\mathtt{s}_3$. Several explanations offer for the light unexpectedly staying off: Relay $\mathtt{re}_1$ might be out of order, resistor $\mathtt{r}_2$ or bulb $\mathtt{li}$ itself might be broken, etc. In order to clarify the situation, a diagnostician may now close switch $\mathtt{s}'_3$. Suppose this activates light bulb $\mathtt{li}_3$, which shows that switch $\mathtt{s}_3$ must have become closed beforehand and also that resistor $\mathtt{r}_3$ is in order. Hence the only remaining diagnosis for the encountered abnormal behavior of the system is a malfunction of light bulb $\mathtt{li}$.

Our example system involves components which are not directly operated, such as the relays and light bulbs. Their state, however, depends in a particular way on the states of other components. These dependences can be expressed in a logical fashion: So-called *state constraints* are logical formulas that constrain the set of potential system states to those which respect the laws of physics. An example constraint for our electric circuit is

$$\mathtt{active}(\mathtt{re}_1) \equiv \mathtt{closed}(\mathtt{s}_1) \tag{1}$$

stating that relay $\mathtt{re}_1$ is activated if and only if switch $\mathtt{s}_1$ is closed. Likewise, the constraint

$$\mathtt{active}(\mathtt{re}_1) \supset \mathtt{closed}(\mathtt{s}_2) \tag{2}$$

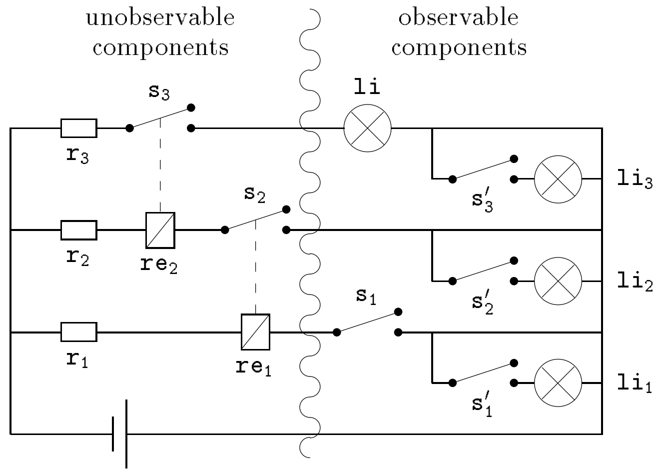

Figure 1: An electric circuit which consists of a number of binary switches; two relays, each of which, in case of activation, attracts the switch located above; three resistors, each needed to keep low the current flow through the respective sub-circuit; and a couple of light bulbs. It is assumed that only the components to the right of the wavelike line are directly observable.

---

formalizes the attracting of switch $\mathtt{s}_2$ by relay $\mathtt{re}_1$ if the latter gets activated. Generally, state constraints are first-order formulas composed of atoms, such as $\mathtt{active}(\mathtt{re}_1)$ etc., which in turn are relations over entities, such as $\mathtt{re}_1$, $\mathtt{s}_1$, etc. Whether or not a particular such atom holds may vary from time to time, as a result of performing actions. Following standard terminology, these atoms are therefore called *fluents*. The truth values of all fluents at a particular point of time determine the current state of the system.

State constraints as means to describe a system is common also in 'static' diagnosis (e.g., [Reiter, 1987]). Crucial for dynamic diagnosis is the observation that state constraints give rise to indirect effects of actions. Closing switch $\mathtt{s}_1$, for instance, has the only immediate effect of $\mathtt{closed}(\mathtt{s}_1)$ becoming true. This, however, causes $\mathtt{active}(\mathtt{re}_1)$ according to (1), which in turn implies that $\mathtt{closed}(\mathtt{s}_2)$ according to (2), etc. In Action Theory, the necessity to account for additional effect of actions which derive from state constraints, is called the *Ramification Problem* [Ginsberg and Smith, 1988a]. Any satisfactory solution requires the successful treatment of two major issues. First, an appropriately weakened version of the commonsense law of persistence needs to be developed which applies only to those parts of the world description that are unaffected by the action's direct *and* indirect effects.[1] Second, not all logical consequences of state constraints may indeed occur as indirect effects [Lifschitz, 1990].[2]

In this paper, we accommodate indirect effects by employing so-called *causal relationships* [Thielscher, 1997b]. These are successively applied subsequent to the generation of the direct effects of an action. For example, this causal relationship is a consequence of state constraint (1) from above:

$$\mathtt{closed(s_1)}\ \ \underline{\text{causes}}\ \ \mathtt{active(re_1)} \tag{3}$$

Informally speaking, it means that whenever $\mathtt{closed(s_1)}$ occurs as direct or indirect effect of an action, then this has the additional, indirect effect $\mathtt{active(re_1)}$. A formal introduction to the theory of causal relationships is given in Section 3. The need for causal relationships in addition to state constraints for system description is due to the dynamic aspect of diagnosis.

Validity of state constraints like (1) and (2), however, relies on the functioning of the involved components, and of course so does the occurrence of the corresponding indirect effects. Our example circuit, for instance, may exhibit a state where, say, switch $\mathtt{s_1}$ is closed but nonetheless relay $\mathtt{re_1}$ remains deactivated. In this case either of the involved components, resistor $\mathtt{r_1}$ or relay $\mathtt{re_1}$, is malfunctioning. In order to reflect this, constraint (1) should actually read weaker—to the effect that

$$\neg\mathtt{ab(r_1)} \wedge \neg\mathtt{ab(re_1)} \supset [\,\mathtt{active(re_1)} \equiv \mathtt{closed(s_1)}\,]$$

where $\mathtt{ab(r_1)}$ and $\mathtt{ab(re_1)}$ represent an abnormal failure of resistor $\mathtt{r_1}$ and relay $\mathtt{re_1}$, respectively. This weakening transfers to our expectations regarding indirect effects: Causal relationship (3) should now read

$$\mathtt{closed(s_1)}\ \ \underline{\text{causes}}\ \ \mathtt{active(re_1)}\ \ \underline{\text{if}}\ \ \neg\mathtt{ab(r_1)} \wedge \neg\mathtt{ab(re_1)}$$

That is to say, closing switch $\mathtt{s_1}$ is expected to causing $\mathtt{active(re_1)}$ only if both the adjacent resistor and the relay itself exhibit their regular behavior.

Under normal circumstances, no system components should malfunction. A diagnosis problem arises as soon as the actual observations contradict the basic assumption that $\bigwedge_{c \in \mathit{components}} \neg\mathtt{ab}(c)$ hold all the time. Diagnosing then amounts to finding one or more affirmative instances of $\mathtt{ab}$ which entail the observed irregular behavior of the system.

Generally, there will be more than a unique collection of affirmative $\mathtt{ab}$-instances that offer as explanation. Telling 'good' from

[1] The commonsense law of persistence says that no system component changes its state when an action is performed unless this change is explicitly mentioned as an effect of that action.

[2] See Section 3.2 for an example.

'bad' diagnoses is a key issue, for the primary diagnosis goal is to find the most likely explanation for the encountered failures. A fundamental principle to this end is minimality: Whenever it suffices to assume that a particular collection of components are malfunctioning, then diagnoses are usually inadequate which assume simultaneous failure of these and other components.[3] Another important issue in view of a good diagnosis is to take into account *a priori* knowledge of differences in the likelihood of components to break. Both these two aspects are standard in diagnosis. Dynamic diagnosis, however, raises an additional challenge when it comes to distinguishing the most plausible diagnoses in case abnormalities are causally connected. The phenomenon of causality naturally arises when dealing with evolutions of systems in the course of time. The challenge is actually more general: It requires to account for a hitherto unnoticed, fundamental problem in Action Theory when dealing with exceptions to ramifications. This will be elucidated in the following section, and a major achievement of this paper is that the resulting theory of actions meets this challenge.

2 The Problem of Causality—and a Solution

In the course of the introduction we have added conditions of 'normality' to both state constraints and the corresponding causal relationships. The intention of doing so was twofold: First, it allows to accommodate situations where the system does not exhibit its regular behavior due to the malfunctioning of components. Second, it supports the search for reasonable diagnoses: Following the principle of minimality, good (i.e., plausible) diagnoses are obtained through suitable *minimization* of abnormality, which means to accept as few instances of `ab` as possible while accounting for the actual observations. This principle shall be illustrated on the basis of the following extract of a system description for our example circuit of the preceding section:

$$
\begin{array}{l}
\neg\mathtt{ab}(\mathtt{r}_1) \wedge \neg\mathtt{ab}(\mathtt{re}_1) \supset [\,\mathtt{active}(\mathtt{re}_1) \equiv \mathtt{closed}(\mathtt{s}_1)\,] \\
\neg\mathtt{ab}(\mathtt{r}_2) \wedge \neg\mathtt{ab}(\mathtt{re}_2) \supset [\,\mathtt{active}(\mathtt{re}_2) \equiv \mathtt{closed}(\mathtt{s}_2)\,] \\
\mathtt{active}(\mathtt{re}_1) \supset \mathtt{closed}(\mathtt{s}_2) \\
\mathtt{active}(\mathtt{re}_2) \supset \mathtt{closed}(\mathtt{s}_3) \\
\neg\mathtt{ab}(\mathtt{li}) \supset [\,\mathtt{active}(\mathtt{li}) \equiv \mathtt{closed}(\mathtt{s}_3)\,]
\end{array}
\qquad (4)
$$

[3] Of course this applies only if component failures are *a priori* rather unlikely. We consider this a fundamental property of the diagnostic problem.

along with some of the corresponding causal relationships, viz.

$$\begin{array}{rlll} \texttt{closed}(\texttt{s}_1) & \underline{\text{causes}} & \texttt{active}(\texttt{re}_1) & \underline{\text{if}} \;\; \neg\texttt{ab}(\texttt{r}_1) \land \neg\texttt{ab}(\texttt{re}_1) \\ \texttt{closed}(\texttt{s}_2) & \underline{\text{causes}} & \texttt{active}(\texttt{re}_2) & \underline{\text{if}} \;\; \neg\texttt{ab}(\texttt{r}_2) \land \neg\texttt{ab}(\texttt{re}_2) \\ \texttt{active}(\texttt{re}_1) & \underline{\text{causes}} & \texttt{closed}(\texttt{s}_2) & \\ \texttt{active}(\texttt{re}_2) & \underline{\text{causes}} & \texttt{closed}(\texttt{s}_3) & \\ \texttt{closed}(\texttt{s}_3) & \underline{\text{causes}} & \texttt{active}(\texttt{li}) & \underline{\text{if}} \;\; \neg\texttt{ab}(\texttt{li}) \end{array} \tag{5}$$

Recall the situation discussed in the introduction, where switch $\texttt{s}_1$ got closed in the state depicted in Figure 1. If light $\texttt{li}$ stays off, then at least one component is out of order. For assuming $\forall x. \neg\texttt{ab}(x)$ in conjunction with the action's effect, $\texttt{closed}(\texttt{s}_1)$, contradicts the observation $\neg\texttt{active}(\texttt{li})$, given the state constraints of equation (4). (This can be seen by the following chain of deductions: $\forall x. \neg\texttt{ab}(x) \land \texttt{closed}(\texttt{s}_1) \land (4) \;\Rightarrow\; \texttt{active}(\texttt{re}_1) \;\Rightarrow\; \texttt{closed}(\texttt{s}_2) \;\Rightarrow\; \texttt{active}(\texttt{re}_2) \;\Rightarrow\; \texttt{closed}(\texttt{s}_3) \;\Rightarrow\; \texttt{active}(\texttt{li})$.) Now, there are five ways of minimizing $\texttt{ab}$ wrt. the formula $\texttt{closed}(\texttt{s}_1) \land \neg\texttt{active}(\texttt{li}) \land (4)$, namely,

$$\begin{array}{rcl} d_1 & = & \{\texttt{ab}(\texttt{r}_1), \neg\texttt{ab}(\texttt{re}_1), \neg\texttt{ab}(\texttt{r}_2), \neg\texttt{ab}(\texttt{re}_2), \neg\texttt{ab}(\texttt{li})\} \\ d_2 & = & \{\neg\texttt{ab}(\texttt{r}_1), \texttt{ab}(\texttt{re}_1), \neg\texttt{ab}(\texttt{r}_2), \neg\texttt{ab}(\texttt{re}_2), \neg\texttt{ab}(\texttt{li})\} \\ d_3 & = & \{\neg\texttt{ab}(\texttt{r}_1), \neg\texttt{ab}(\texttt{re}_1), \texttt{ab}(\texttt{r}_2), \neg\texttt{ab}(\texttt{re}_2), \neg\texttt{ab}(\texttt{li})\} \\ d_4 & = & \{\neg\texttt{ab}(\texttt{r}_1), \neg\texttt{ab}(\texttt{re}_1), \neg\texttt{ab}(\texttt{r}_2), \texttt{ab}(\texttt{re}_2), \neg\texttt{ab}(\texttt{li})\} \\ d_5 & = & \{\neg\texttt{ab}(\texttt{r}_1), \neg\texttt{ab}(\texttt{re}_1), \neg\texttt{ab}(\texttt{r}_2), \neg\texttt{ab}(\texttt{re}_2), \texttt{ab}(\texttt{li})\} \end{array} \tag{6}$$

If, for the sake of simplicity, we assume for the moment that failures of resistors, relays, and light bulbs are equally likely, then $d_1, \ldots, d_5$ together are the five diagnoses which are reasonably to be expected here. Far less plausible would be, say, the diagnosis that simultaneously the two relays and also the bulb are malfunctioning. Minimizing abnormality thus determines exactly the plausible diagnoses in this example. This is a well-established result as far as static diagnosis is concerned, where abnormalities are causally independent (see, e.g., [Reiter, 1987]).

Unfortunately, however, this standard way of minimizing abnormality turns out problematic as soon as causal interactions among abnormalities need to be taken into account. This shall be illustrated by the following scenario. Let us add to our system description the knowledge that in our example circuit a relay gets broken whenever it forms an active sub-circuit with a broken resistor. This is represented by these two additional state constraints:

$$\begin{array}{l} \texttt{ab}(\texttt{r}_1) \land \texttt{closed}(\texttt{s}_1) \supset \texttt{ab}(\texttt{re}_1) \\ \texttt{ab}(\texttt{r}_2) \land \texttt{closed}(\texttt{s}_2) \supset \texttt{ab}(\texttt{re}_2) \end{array} \tag{7}$$

They give rise to indirect effects as follows:

$$\begin{array}{llll} \texttt{closed}(\texttt{s}_1) & \underline{\text{causes}} & \texttt{ab}(\texttt{re}_1) & \underline{\text{if}} \;\; \texttt{ab}(\texttt{r}_1) \\ \texttt{closed}(\texttt{s}_2) & \underline{\text{causes}} & \texttt{ab}(\texttt{re}_2) & \underline{\text{if}} \;\; \texttt{ab}(\texttt{r}_2) \end{array} \tag{8}$$

That is to say, as soon as the respective sub-circuit with the broken resistor gets closed the relay breaks as well. Thus the abnormalities $\mathtt{ab}(\mathtt{r}_i)$ and $\mathtt{ab}(\mathtt{re}_i)$ (for $i = 1, 2$) become causally connected.

To see how the introduction of causal dependencies among abnormalities affects minimization, suppose we already know that in the situation depicted in Figure 1 resistor $\mathtt{r}_2$ is broken, i.e., that $\mathtt{ab}(\mathtt{r}_2)$ holds. What effect is to be expected when closing switch $\mathtt{s}_1$? Since nothing hints at either resistor $\mathtt{r}_1$ or relay $\mathtt{re}_1$ malfunctioning, we should expect that $\mathtt{re}_1$ is activated and will thus attract switch $\mathtt{s}_2$. This switch getting closed in turn will cause relay $\mathtt{re}_2$ to break according to (8), given that $\mathtt{ab}(\mathtt{r}_2)$. Hence, one intuitively expects that the effect $\mathtt{ab}(\mathtt{re}_2)$ materialize.

But what happens if abnormality is globally minimized in this scenario? It is clear that one additional abnormality aside from the given $\mathtt{ab}(\mathtt{r}_2)$ is inevitable. Formally, this follows from $\mathtt{ab}(\mathtt{r}_2) \wedge \mathtt{closed}(\mathtt{s}_1) \wedge (4) \wedge (7)$ being inconsistent with the assumption that $\neg\mathtt{ab}(c)$ holds for each $c \neq \mathtt{r}_2$. Therefore, one minimal model reflects the above conclusion that $\mathtt{ab}(\mathtt{re}_2)$. This corresponds to the intended model. Yet abnormality can be minimized in more ways. Namely, we can try to assume an exception to the very first ramification, i.e., the one which activates relay $\mathtt{re}_1$. This assumption requires to grant that either $\mathtt{ab}(\mathtt{re}_1)$ or $\mathtt{ab}(\mathtt{r}_1)$ hold. But for compensation, now that relay $\mathtt{re}_1$ does not get activated we avoid the conclusion that switch $\mathtt{s}_2$ gets closed, hence that relay $\mathtt{re}_2$ breaks. In other words, accepting $\mathtt{ab}(\mathtt{r}_1)$ or $\mathtt{ab}(\mathtt{re}_1)$ allows to assume $\neg\mathtt{ab}(\mathtt{re}_2)$. We thus obtain a second and third minimal model here, which in total gives us these three suggested outcomes:

$$
\begin{aligned}
d_1 &= \{\neg\mathtt{ab}(\mathtt{r}_1), \neg\mathtt{ab}(\mathtt{re}_1), \mathtt{ab}(\mathtt{r}_2), \mathtt{ab}(\mathtt{re}_2), \neg\mathtt{ab}(\mathtt{li})\} \\
d_2 &= \{\mathtt{ab}(\mathtt{r}_1), \neg\mathtt{ab}(\mathtt{re}_1), \mathtt{ab}(\mathtt{r}_2), \neg\mathtt{ab}(\mathtt{re}_2), \neg\mathtt{ab}(\mathtt{li})\} \\
d_3 &= \{\neg\mathtt{ab}(\mathtt{r}_1), \mathtt{ab}(\mathtt{re}_1), \mathtt{ab}(\mathtt{r}_2), \neg\mathtt{ab}(\mathtt{re}_2), \neg\mathtt{ab}(\mathtt{li})\}
\end{aligned}
$$

Each of d_1, d_2, d_3 minimizes abnormality, but only d_1 entails the expected effect $\mathtt{ab}(\mathtt{re}_2)$. The two additional models are therefore unintended.

To stress the point, both d_2 and d_3 should even be called counter-intuitive, and this is not because an abnormality of relay $\mathtt{re}_2$ is *a priori* more likely than an abnormality of resistor $\mathtt{r}_1$ or of relay $\mathtt{re}_1$. On the contrary: Model d_1 should be preferred even if, say, $\mathtt{ab}(\mathtt{r}_1)$ had a higher prior likelihood than $\mathtt{ab}(\mathtt{re}_2)$. For what decisively distinguishes d_1 from both d_2 and d_3 is that $\mathtt{ab}(\mathtt{re}_2)$ but neither $\mathtt{ab}(\mathtt{r}_1)$ nor $\mathtt{ab}(\mathtt{re}_1)$ can easily be explained from the perspective of causality in this particular situation: Closing switch $\mathtt{s}_1$ along with all of its expected indirect effects *causes* the fact that $\mathtt{ab}(\mathtt{re}_2)$ finally holds, whereas $\mathtt{ab}(\mathtt{r}_1)$ and $\mathtt{ab}(\mathtt{re}_1)$ come out of the blue in the unintended minimal models.

This disturbing observation resembles a key problem in the context of the Qualification Problem in reasoning about actions [Mc-

Carthy, 1977] if the latter is approached without supporting the distinction between caused and unmotivated disqualifications of actions [Lifschitz, 1987].[4] The reader may also notice the similarities to the famous Yale Shooting counter-example [Hanks and McDermott, 1987]: A gun that becomes magically unloaded while waiting truly deserves being called abnormal, whereas causality explains the death of the turkey if being shot at with a loaded gun.

The key to a solution is to respect causality by conducting the minimization step at the right time. Notice that the unintended models d_2 and d_3 have been obtained by minimizing $\mathtt{ab}$ in the *resulting* state (as has d_1). This did not allow for taking into account the crucial causal dependence, for the phenomenon of causality manifests in state transitions but not in a single, static state. The alternative is to concentrate on the *initial* state when minimizing $\mathtt{ab}$, i.e., on the state prior to the closing of switch $\mathtt{s}_1$.

Suppose again given $\mathtt{ab}(\mathtt{r}_2)$, but now switch $\mathtt{s}_1$ shall still be open. Then it is consistent to assume that all other instances of $\mathtt{ab}$ are false. More precisely, $\mathtt{ab}(\mathtt{r}_2) \wedge \neg\mathtt{closed}(\mathtt{s}_1) \wedge (4) \wedge (7)$ admits a unique $\mathtt{ab}$-minimal model, viz.

$$d_0 \;=\; \{\neg\mathtt{ab}(\mathtt{r}_1), \neg\mathtt{ab}(\mathtt{re}_1), \mathtt{ab}(\mathtt{r}_2), \neg\mathtt{ab}(\mathtt{re}_2), \neg\mathtt{ab}(\mathtt{li})\}$$

Now, if switch $\mathtt{s}_1$ is closed in the state which is depicted in Figure 1 and which satisfies d_0, then the only possible resulting state satisfies d_1, as intended. In particular, causality 'naturally' brings about the additional abnormality $\mathtt{ab}(\mathtt{re}_2)$ as indirect effect: According to the topmost causal relationship in (5), $\mathtt{closed}(\mathtt{s}_1)$ causes $\mathtt{active}(\mathtt{re}_1)$ given that $\neg\mathtt{ab}(\mathtt{r}_1) \wedge \neg\mathtt{ab}(\mathtt{re}_1)$. This in turn causes $\mathtt{closed}(\mathtt{s}_2)$ following the third causal relationship in (5). After that, finally, the second causal relationship in (8) becomes applicable, yielding $\mathtt{ab}(\mathtt{re}_2)$. The additional, *caused* abnormality is thus accounted for by means of ramification—and not by means of minimization.

In case one has to deal with a whole sequence of actions, the above argument needs to be iterated. If minimizing abnormality in the finally resulting state risks to ignore causal information, then so does minimization conducted in the final but one state, and so on. Consequently, when the diagnostician reasons about the actions that have been taken, then he or she should perform the minimization step

[4] An example is the *Berkeley Rascal Trap* [Thielscher, 1996a]: Suppose that the action of inserting a potato into the tail pipe of a car is abnormally disqualified if the potato is too heavy, and that the action of starting the engine of the car is abnormally disqualified if the tail pipe houses a potato. Then we should expect difficulties with starting the engine if a little rascal first tried to put a potato into the tail pipe. But globally minimizing abnormalities in this example produces a second model where the action of introducing a potato is disqualified in the first place. While this disqualification is to be considered abnormal, it avoids a disqualification of the following action of starting the engine. Thus this second model minimizes abnormality as well, though it is obviously counter-intuitive.

as early as possible in order to exploit as much as possible causal information. This minimizing initially is justified by the commonsense assumption that causality is effective only forward in time, by which it is clear that no causal reason for an abnormality in the initial state can possibly be known of. Of course this does not imply that such a causal reason does not exist. But if it does, then it is not part of the diagnostician's knowledge, hence has no influence on the correct reasoning about this knowledge. The general paradigm of initial minimization has previously been successfully employed for reasoning about space occupancy [Shanahan, 1995], for minimizing events in narratives [Thielscher, 1998], and to account for causality in the context of the Qualification Problem [Thielscher, 1996a].

In the following but one section, we present a formal theory of dynamic diagnosis which reflects the insights gained in this section. Prior to this, we recall from [Thielscher, 1997b] the formal notions and notations related to the theory of causal relationships as means to solve the plain Ramification Problem.

3 Causal Relationships and the Ramification Problem

In formal systems for reasoning about actions, the Ramification Problem denotes the problem of handling indirect effects. As such, these effects are not explicitly represented in action specifications but follow from general domain knowledge, formalized as state constraints. Recent research has revealed that incorporating the commonsense notion of causality helps with solving this problem (e.g., [Elkan, 1992, Lin, 1995, McCain and Turner, 1995, Thielscher, 1997b]). The theory of causal relationships provides an approach along this line. In this section we repeat the formal definitions underlying this theory. Our goal is to provide a formalism which allows us to specify the behavior of dynamic systems in terms of direct and indirect effects of actions. We then take the resulting formalism as the basis for a theory of dynamic diagnosis.

3.1 Fluents and States, Actions and Change

The concept of a *state* is fundamental for dealing with dynamic systems. A state is a snapshot of the system being modeled at a particular instant of time. States are composed of atomic propositions, so-called *fluents*, which represent properties of *entities*. The truth-value of any such proposition may change in the course of time as a consequence of state transition, and each state is characterized by a particular combination of truth values of all fluents.

Definition 1 Let $\mathcal{E}$ be a finite set of symbols called *entities*. Let $\mathcal{F}$ denote a finite set of symbols called *fluent names*, each of which is associated with a natural number (the *arity*) and a *scope*, indicating which entities may serve as arguments.

A *fluent* is an expression $f(e_1, \ldots, e_n)$ where $f \in \mathcal{F}$ is of arity n and where the n-tuple $(e_1, \ldots, e_n) \in \mathcal{E}^n$ belongs to the scope of f. A *fluent literal* is a fluent or its negation $\neg f(e_1, \ldots, e_n)$. A set of fluent literals is *inconsistent* if it contains a fluent along with its negation, otherwise it is *consistent*. A *state* is a maximal consistent set of fluent literals. ■

Example 1 Our electric circuit consists of the following 15 entities (6 switches, 4 bulbs, 3 resistors, and 2 relays):

$$\mathcal{E} = \{\mathtt{s}_1, \mathtt{s}_1', \mathtt{s}_2, \mathtt{s}_2', \mathtt{s}_3, \mathtt{s}_3', \mathtt{li}, \mathtt{li}_1, \mathtt{li}_2, \mathtt{li}_3, \mathtt{r}_1, \mathtt{r}_2, \mathtt{r}_3, \mathtt{re}_1, \mathtt{re}_2\}$$

The various states our circuit may exhibit shall be described using the three unary fluent names `closed`, `active`, and `ab`. The first ranges over all switches, the scope of the second are both light bulbs and relays, and the scope of the third are bulbs, relays, and resistors. Examples for fluents are $\mathtt{closed}(\mathtt{s}_2')$, $\mathtt{active}(\mathtt{li})$, and $\mathtt{ab}(\mathtt{r}_1)$, but not, say, $\mathtt{closed}(\mathtt{re}_1)$ or $\mathtt{ab}(\mathtt{s}_2')$. In this way, the system state depicted in Figure 1 reads as follows if we assume that all components are in order:

$$\begin{aligned} S_0 = \{ & \neg\mathtt{closed}(\mathtt{s}_1), \neg\mathtt{closed}(\mathtt{s}_1'), \ldots, \\ & \neg\mathtt{active}(\mathtt{li}), \ldots, \neg\mathtt{active}(\mathtt{re}_1), \neg\mathtt{active}(\mathtt{re}_2), \\ & \neg\mathtt{ab}(\mathtt{li}), \ldots, \neg\mathtt{ab}(\mathtt{r}_1), \neg\mathtt{ab}(\mathtt{r}_2), \neg\mathtt{ab}(\mathtt{re}_1), \neg\mathtt{ab}(\mathtt{re}_2) \} \end{aligned} \tag{9}$$

■

For convenience, we will use the following notational conventions: If ℓ is a fluent literal, then by $\|\ell\|$ we denote its affirmative component, that is, $\|f(\widetilde{e})\| = \|\neg f(\widetilde{e})\| = f(\widetilde{e})$ where $f \in \mathcal{F}$ and $\widetilde{e}$ is a sequence of n entities with n being the arity of f. This notation extends to sets S of fluent literals as follows: $\|S\| = \{\|\ell\| : \ell \in S\}$. E.g., whenever S is a state, then $\|S\|$ is the set of all fluents. If ℓ is a negative fluent literal, then $\neg\ell$ should be interpreted as $\|\ell\|$. In other words, $\neg\neg f(\widetilde{e}) = f(\widetilde{e})$. Finally, if S is a set of fluent literals, then by $\neg S$ we denote the set $\{\neg\ell : \ell \in S\}$. E.g., a set S of fluent literals is inconsistent iff $S \cap \neg S \neq \{\}$.

The elements of an underlying set of fluents can be considered atoms for constructing formulas using the standard logical connectives. Fluent formulas are needed to describe dependences among the state components. Formally, each fluent literal (possibly containing variables[5]) is considered a fluent formulas; $\top$ (tautology) and $\bot$ (contradiction) are fluent formulas; and if F and G are fluent formulas, then so are $\neg F$, $F \wedge G$, $F \vee G$, $F \supset G$, $F \equiv G$, $\exists x. F$, and

[5] Since the argument space of a fluent may be restricted according to a designated scope, we formally need to attach sorts to variables. In what follows, it is assumed that fluent literals with variables are always well-formed in that the scope of the fluent is respected by the sorts of these variables.

$$
\begin{array}{rcl}
\neg\mathtt{ab}(\mathtt{r}_1) \wedge \neg\mathtt{ab}(\mathtt{re}_1) & \supset & [\ \mathtt{active}(\mathtt{re}_1) \equiv \mathtt{closed}(\mathtt{s}_1)\ ] \\
\neg\mathtt{ab}(\mathtt{r}_2) \wedge \neg\mathtt{ab}(\mathtt{re}_2) & \supset & [\ \mathtt{active}(\mathtt{re}_2) \equiv \mathtt{closed}(\mathtt{s}_2)\ ] \\
\neg\mathtt{ab}(\mathtt{r}_3) \wedge \neg\mathtt{ab}(\mathtt{li}) & \supset & [\ \mathtt{active}(\mathtt{li}) \equiv \mathtt{closed}(\mathtt{s}_3)\ ] \\
\neg\mathtt{ab}(\mathtt{r}_1) \wedge \neg\mathtt{ab}(\mathtt{li}_1) & \supset & [\ \mathtt{active}(\mathtt{li}_1) \equiv \mathtt{closed}(\mathtt{s}_1) \wedge \mathtt{closed}(\mathtt{s}_1')\ ] \\
\neg\mathtt{ab}(\mathtt{r}_2) \wedge \neg\mathtt{ab}(\mathtt{li}_2) & \supset & [\ \mathtt{active}(\mathtt{li}_2) \equiv \mathtt{closed}(\mathtt{s}_2) \wedge \mathtt{closed}(\mathtt{s}_2')\ ] \\
\neg\mathtt{ab}(\mathtt{r}_3) \wedge \neg\mathtt{ab}(\mathtt{li}_3) & \supset & [\ \mathtt{active}(\mathtt{li}_3) \equiv \mathtt{closed}(\mathtt{s}_3) \wedge \mathtt{closed}(\mathtt{s}_3')\ ] \\
\mathtt{active}(\mathtt{re}_1) & \supset & \mathtt{closed}(\mathtt{s}_2) \\
\mathtt{active}(\mathtt{re}_2) & \supset & \mathtt{closed}(\mathtt{s}_3) \\
\forall x\ [\ \mathtt{ab}(x) & \supset & \neg\mathtt{active}(x)\ ] \\
\mathtt{ab}(\mathtt{r}_1) \wedge \mathtt{closed}(\mathtt{s}_1) & \supset & \mathtt{ab}(\mathtt{re}_1) \\
\mathtt{ab}(\mathtt{r}_2) \wedge \mathtt{closed}(\mathtt{s}_2) & \supset & \mathtt{ab}(\mathtt{re}_2)
\end{array}
$$

Figure 2: All state constraints for our example circuit.

$\forall x.\, F$ (where x is a variable). A *closed* formula is a fluent formula without free variables. The notion of closed fluent formulas being *true* in a state S is inductively defined as usual:

1. $\top$ is true and $\bot$ is false in S;
2. a fluent literal ℓ is true in S iff $\ell \in S$;
3. $F \wedge G$ is true in S iff F and G are true in S;
4. $F \vee G$ is true in S iff F or G is true in S (or both);
5. $F \supset G$ is true in S iff F is false in S or G is true in S (or both);
6. $F \equiv G$ is true in S iff F and G are true in S, or else F and G are false in S;
7. $\exists x.\, F$ is true in S iff there exists some $e \in \mathcal{E}$ such that $F\{x \mapsto e\}$ is true in S;
8. $\forall x.\, F$ is true in S iff for each $e \in \mathcal{E}$, $F\{x \mapsto e\}$ is true in S.

Here, $F\{x \mapsto e\}$ denotes the fluent formula resulting from replacing in F all free occurrences of x by entity e (which should belong to the sort of x; c.f. footnote 5). *State constraints* are closed fluent formulas which all physically possible states of a system satisfy. These states are also called *acceptable*.

Example 1 (continued) The 11 state constraints depicted in Figure 2 describe all the various physical relations among the components in our example circuit. These constraints hold in S_0 as defined in equation (9), but they are violated, for instance, in a state where $\mathtt{closed}(\mathtt{s}_1)$, $\forall x.\, \neg\mathtt{ab}(x)$, and $\neg\mathtt{active}(\mathtt{li})$ hold. ■

The second fundamental notion in Action Theory are the actions themselves. Actions cause state transitions. The (direct) effect of an action is specified by saying which fluents change their truth-value when the action is being performed. The formal notion of *action laws* serves this purpose.

Definition 2 Let $\mathcal{A}$ be a finite set of action names, each of which is associated with an arity and a scope. An *action* is a ground term $a(e_1, \ldots, e_n)$ where $a \in \mathcal{A}$ is of arity n and the n-tuple $(e_1, \ldots, e_n) \in \mathcal{E}^n$ belongs to the scope of a.

An *action law* is of the form

$$a(\widetilde{x}) \ \underline{\text{transforms}} \ C \ \underline{\text{into}} \ E$$

where $\widetilde{x}$ is a sequence of pairwise distinct variables, $a \in \mathcal{A}$ is of arity equal to the length of $\widetilde{x}$, and where C (the *condition*) and E (the *effect*) are sets of fluent literals (possibly with variables chosen from $\widetilde{x}$) which satisfy the following. Both $C\{\widetilde{x} \mapsto \widetilde{e}\}$ and $E\{\widetilde{x} \mapsto \widetilde{e}\}$, for any sequence of entities $\widetilde{e}$ in the scope of a, are consistent; and moreover, $\|C\{\widetilde{x} \mapsto \widetilde{e}\}\| = \|E\{\widetilde{x} \mapsto \widetilde{e}\}\|$, that is, condition and effect always refer to the same fluents.[6] If S is a state, then a ground instance $\alpha\{\widetilde{x} \mapsto \widetilde{e}\}$ of an action law $\alpha = a(\widetilde{x})$ $\underline{\text{transforms}}$ C $\underline{\text{into}}$ E is *applicable* to S iff $C\{\widetilde{x} \mapsto \widetilde{e}\} \subseteq S$. The *application* of $\alpha\{\widetilde{x} \mapsto \widetilde{e}\}$ to S yields $(S \setminus C\{\widetilde{x} \mapsto \widetilde{e}\}) \cup E\{\widetilde{x} \mapsto \widetilde{e}\}$. ■

Notably, due to $\|C\| = \|E\|$ the resulting set $(S \setminus C) \cup E$ is a state if so is S, but it may violate the underlying state constraints. As an example, consider the action of toggling a switch. We use the unary action name `toggle` in conjunction with these two action laws:

$$\begin{array}{llll} \mathtt{toggle}(x) & \underline{\text{transforms}} & \{\neg\mathtt{closed}(x)\} & \underline{\text{into}} \ \{\mathtt{closed}(x)\} \\ \mathtt{toggle}(x) & \underline{\text{transforms}} & \{\mathtt{closed}(x)\} & \underline{\text{into}} \ \{\neg\mathtt{closed}(x)\} \end{array} \tag{10}$$

If we perform the action $\mathtt{toggle(s_1)}$ in state S_0 from above (c.f. equation (9)), then the first of the two laws is applicable on account of $\{\neg\mathtt{closed(s_1)}\} \subseteq S_0$. The resulting state is

$$\begin{aligned} S_1 = \{ & \mathtt{closed(s_1)}, \neg\mathtt{closed(s_1')}, \ldots, \\ & \neg\mathtt{active(li)}, \ldots, \neg\mathtt{active(re_1)}, \neg\mathtt{active(re_2)}, \\ & \neg\mathtt{ab(li)}, \ldots, \neg\mathtt{ab(r_1)}, \neg\mathtt{ab(r_2)}, \neg\mathtt{ab(re_1)}, \neg\mathtt{ab(re_2)} \} \end{aligned} \tag{11}$$

This state violates the state constraints of Figure 2 since only the immediate effect of toggling switch $\mathtt{s_1}$ has been computed.

Our Definition 2 does not exclude the existence of two or more simultaneously applicable laws for one and the same action. This supports the specification of actions with indeterminate effects, so-called non-deterministic actions. Suppose, for example, it is totally

[6] If $\tilde{x} = x_1, \ldots x_n$ and $\tilde{e} = e_1, \ldots, e_n$, then $\{\tilde{x} \mapsto \tilde{e}\}$ means the simultaneous replacing $\{x_1 \mapsto e_1\}, \ldots, \{x_n \mapsto e_n\}$. The requirement that condition and effect concern the very same fluents simplifies the definition of how action laws are applied. It does not impose a restriction of expressiveness since we allow several laws for the same action.

dark so that it is impossible to tell apart the three switches $\mathtt{s}'_1$, $\mathtt{s}'_2$, and $\mathtt{s}'_3$. Nonetheless we want to close one of them (knowing they all are currently open). Putting this plan into execution, there are three possible outcomes: We either hit the first, the second, or else the third switch. This may be formalized by the three action laws

$$\mathtt{close\text{-}a\text{-}switch}\ \ \underline{\text{transforms}}\ \ \{\neg\mathtt{closed}(\mathtt{s}'_1)\}\ \ \underline{\text{into}}\ \ \{\mathtt{closed}(\mathtt{s}'_1)\}$$
$$\mathtt{close\text{-}a\text{-}switch}\ \ \underline{\text{transforms}}\ \ \{\neg\mathtt{closed}(\mathtt{s}'_2)\}\ \ \underline{\text{into}}\ \ \{\mathtt{closed}(\mathtt{s}'_2)\}$$
$$\mathtt{close\text{-}a\text{-}switch}\ \ \underline{\text{transforms}}\ \ \{\neg\mathtt{closed}(\mathtt{s}'_3)\}\ \ \underline{\text{into}}\ \ \{\mathtt{closed}(\mathtt{s}'_3)\}$$

Each one of these laws is applicable to a state where all three switches are open, and they yield different resulting states when being applied.

3.2 Indirect effects

In all but very simple environments actions usually have greater impact than what is specified in action laws. These laws describe the *direct* effects of actions. Toggling a switch, for instance, has the only direct effect of that very switch changing its position. The theory of causal relationships takes the stance that the state is merely intermediate which results from accounting just for the direct effect. That state may require further computation to accommodate additional, *indirect* effects. In our running example, possible indirect effects are activations of light bulbs and relays, or the attraction of a switch by some relay.

To be more specific, each single indirect effect is obtained according to so-called causal relationships, whose formal definition is as follows.

Definition 3 Let $\mathcal{E}$ and $\mathcal{F}$ be sets of entities and fluent names, respectively. A *causal relationship* is of the form

$$\varepsilon\ \ \underline{\text{causes}}\ \ \varrho\ \ \underline{\text{if}}\ \ \Phi$$

where Φ (the *context*) is a fluent formula and both ε (the *triggering effect*) and ϱ (the *ramification*) are fluent literals (possibly containing variables). ■

The intended reading is the following: Under condition Φ, the (previously obtained, direct or indirect) effect ε triggers the indirect effect ϱ. For notational convenience, we use ε <u>causes</u> ϱ as a shorthand form of the causal relationship ε <u>causes</u> ϱ <u>if</u> $\top$.

We have somewhat loosely said that indirect effects are consequences of state constraints. Having the formal definition of causal relationships, this correspondence can be stated more precisely. A causal relationship ε <u>causes</u> ϱ <u>if</u> Φ is consequence of some state constraint if the latter implies $\Phi \wedge \varepsilon \supset \varrho$. However, not all such purely logical consequences of state constraints correspond to indirect effects, as has first been observed in [Lifschitz, 1990]. To see why, recall our state constraint $\texttt{active(li}_1\texttt{)} \equiv \texttt{closed(s}_1\texttt{)} \wedge \texttt{closed(s}'_1\texttt{)}$. Among its logical consequences are the two implications

$$\texttt{closed(s}'_1\texttt{)} \wedge \texttt{closed(s}_1\texttt{)} \supset \texttt{active(li}_1\texttt{)}$$
$$\neg\texttt{active(li}_1\texttt{)} \wedge \texttt{closed(s}_1\texttt{)} \supset \neg\texttt{closed(s}'_1\texttt{)}$$

Yet only the first one gives rise to a valid causal relationship, viz.

$$\texttt{closed(s}_1\texttt{)}\ \underline{\text{causes}}\ \texttt{active(li}_1\texttt{)}\ \underline{\text{if}}\ \texttt{closed(s}'_1\texttt{)}$$

The second of the two implications, if taken as causal relationship, would read

$$\texttt{closed(s}_1\texttt{)}\ \underline{\text{causes}}\ \neg\texttt{closed(s}'_1\texttt{)}\ \underline{\text{if}}\ \neg\texttt{active(li}_1\texttt{)}$$

In other words, closing switch $\texttt{s}_1$ would cause switch $\texttt{s}'_1$ to open rather than light bulb $\texttt{li}_1$ to becoming activated. This is obviously an undesired conclusion. The observation that a state constraint may not contain sufficient information to tell apart its *causal* consequences was the striving force for developing the theory of causal relationships. Causal relationships thus contain more information than the mere state constraints. It is, however, not necessary to draw them up all by hand. Causal relationships can rather be generated automatically given additional domain-specific knowledge—called *influence information*—of how fluents may generally affect each other. For details see [Thielscher, 1997b].

Example 1 (continued) The 23 causal relationships shown in Figure 3 represent all indirect effect that can possibly occur in our example circuit. They derive from the various state constraints listed in Figure 2.[7] ■

The application of a causal relationship yields a single indirect effect. To reiterate this process, causal relationships repeatedly manipulate state-effect pairs (S, E): State S is an intermediate result where some but not yet all indirect effects have been accounted for, and E contains all direct and indirect effects computed so far.

Definition 4 Let (S, E) be a pair consisting of a state S and a set of fluent literals E. Furthermore, let $r = \varepsilon\ \underline{\text{causes}}\ \varrho\ \underline{\text{if}}\ \Phi$ be a causal relationship, and let $\widetilde{x}$ denote a sequence of all free variables occurring in ε, ϱ, or Φ. Then a ground instance $r\{\widetilde{x} \mapsto \widetilde{e}\}$ is *applicable* to (S, E) iff $\varepsilon\{\widetilde{x} \mapsto \widetilde{e}\} \in E$ and $\Phi\{\widetilde{x} \mapsto \widetilde{e}\} \wedge \neg\varrho\{\widetilde{x} \mapsto \widetilde{e}\}$ is true in S. The *application* of $r\{\widetilde{x} \mapsto \widetilde{e}\}$ to (S, E) yields the pair (S', E') where $S' = (S \setminus \{\neg\varrho\{\widetilde{x} \mapsto \widetilde{e}\}\}) \cup \{\varrho\{\widetilde{x} \mapsto \widetilde{e}\}\}$ and $E' = (E \setminus \{\neg\varrho\{\widetilde{x} \mapsto \widetilde{e}\}\}) \cup \{\varrho\{\widetilde{x} \mapsto \widetilde{e}\}\}$. ■

[7] As indicated, these causal relationships can be automatically obtained from our state constraints by providing the additional domain knowledge that changing the position of a switch does have the potential to affect certain light bulbs and relays, and that each relay has the potential to affect the opposite switch's position.

$\mathtt{closed(s_1)}$	causes	$\mathtt{active(re_1)}$	if	$\neg\mathtt{ab(r_1)} \wedge \neg\mathtt{ab(re_1)}$
$\neg\mathtt{closed(s_1)}$	causes	$\neg\mathtt{active(re_1)}$	if	$\neg\mathtt{ab(r_1)} \wedge \neg\mathtt{ab(re_1)}$
$\mathtt{closed(s_2)}$	causes	$\mathtt{active(re_2)}$	if	$\neg\mathtt{ab(r_2)} \wedge \neg\mathtt{ab(re_2)}$
$\neg\mathtt{closed(s_2)}$	causes	$\neg\mathtt{active(re_2)}$	if	$\neg\mathtt{ab(r_2)} \wedge \neg\mathtt{ab(re_2)}$
$\mathtt{closed(s_3)}$	causes	$\mathtt{active(li)}$	if	$\neg\mathtt{ab(r_3)} \wedge \neg\mathtt{ab(li)}$
$\neg\mathtt{closed(s_3)}$	causes	$\neg\mathtt{active(li)}$	if	$\neg\mathtt{ab(r_3)} \wedge \neg\mathtt{ab(li)}$
$\mathtt{closed(s_1)}$	causes	$\mathtt{active(li_1)}$	if	$\mathtt{closed(s_1')} \wedge \neg\mathtt{ab(r_1)} \wedge \neg\mathtt{ab(li_1)}$
$\mathtt{closed(s_1')}$	causes	$\mathtt{active(li_1)}$	if	$\mathtt{closed(s_1)} \wedge \neg\mathtt{ab(r_1)} \wedge \neg\mathtt{ab(li_1)}$
$\neg\mathtt{closed(s_1)}$	causes	$\neg\mathtt{active(li_1)}$	if	$\neg\mathtt{ab(r_1)} \wedge \neg\mathtt{ab(li_1)}$
$\neg\mathtt{closed(s_1')}$	causes	$\neg\mathtt{active(li_1)}$	if	$\neg\mathtt{ab(r_1)} \wedge \neg\mathtt{ab(li_1)}$
$\mathtt{closed(s_2)}$	causes	$\mathtt{active(li_2)}$	if	$\mathtt{closed(s_2')} \wedge \neg\mathtt{ab(r_2)} \wedge \neg\mathtt{ab(li_2)}$
$\mathtt{closed(s_2')}$	causes	$\mathtt{active(li_2)}$	if	$\mathtt{closed(s_2)} \wedge \neg\mathtt{ab(r_2)} \wedge \neg\mathtt{ab(li_2)}$
$\neg\mathtt{closed(s_2)}$	causes	$\neg\mathtt{active(li_2)}$	if	$\neg\mathtt{ab(r_2)} \wedge \neg\mathtt{ab(li_2)}$
$\neg\mathtt{closed(s_2')}$	causes	$\neg\mathtt{active(li_2)}$	if	$\neg\mathtt{ab(r_2)} \wedge \neg\mathtt{ab(li_2)}$
$\mathtt{closed(s_3)}$	causes	$\mathtt{active(li_3)}$	if	$\mathtt{closed(s_3')} \wedge \neg\mathtt{ab(r_3)} \wedge \neg\mathtt{ab(li_3)}$
$\mathtt{closed(s_3')}$	causes	$\mathtt{active(li_3)}$	if	$\mathtt{closed(s_3)} \wedge \neg\mathtt{ab(r_3)} \wedge \neg\mathtt{ab(li_3)}$
$\neg\mathtt{closed(s_3)}$	causes	$\neg\mathtt{active(li_3)}$	if	$\neg\mathtt{ab(r_3)} \wedge \neg\mathtt{ab(li_3)}$
$\neg\mathtt{closed(s_3')}$	causes	$\neg\mathtt{active(li_3)}$	if	$\neg\mathtt{ab(r_3)} \wedge \neg\mathtt{ab(li_3)}$
$\mathtt{active(re_1)}$	causes	$\mathtt{closed(s_2)}$		
$\mathtt{active(re_2)}$	causes	$\mathtt{closed(s_3)}$		
$\mathtt{ab}(x)$	causes	$\neg\mathtt{active}(x)$		
$\mathtt{closed(s_1)}$	causes	$\mathtt{ab(re_1)}$	if	$\mathtt{ab(r_1)}$
$\mathtt{closed(s_2)}$	causes	$\mathtt{ab(re_2)}$	if	$\mathtt{ab(r_2)}$

Figure 3: The causal relationships that hold in our example circuit.

In words, a causal relationship is applicable if the associated condition Φ holds in S, the particular indirect effect ϱ is currently false, and the cause ε is among the current effects, E. As the result of the application the indirect effect ϱ becomes true in S and is added to E. If $\mathcal{R}$ is a set of causal relationships, then by $(S,E) \leadsto_{\mathcal{R}} (S',E')$ we denote the existence of an element in $\mathcal{R}$ whose application to (S,E) yields (S',E'). Notice that if S is a state and E is consistent, then $(S,E) \leadsto_{\mathcal{R}} (S',E')$ implies that S' is a state and E' is consistent, too. We adopt a standard notation in writing $(S,E) \overset{*}{\leadsto}_{\mathcal{R}} (S',E')$ to indicate that there are causal relationships in $\mathcal{R}$ whose successive application to (S,E) yields (S',E').

Now, suppose given a preliminary state S as the result of having computed the direct effect E of an action via Definition 2. Additional, indirect effects are then accounted for by (non-deterministically) selecting and (serially) applying causal relationships until a state satisfying the state constraints obtains.

Definition 5 Let $\mathcal{E}$, $\mathcal{F}$, $\mathcal{A}$, and $\mathcal{L}$ be sets of entities, fluent names, action names, and action laws, respectively. Furthermore, let $\mathcal{C}$ be a set of state constraints and $\mathcal{R}$ a set of causal relationships. If S is an acceptable state and a an action, then a state S' is a *successor state* of S and a iff the following holds: Set $\mathcal{L}$ contains an applicable instance a transforms C into E of an action law such that

1. $((S \setminus C) \cup E, E) \overset{*}{\leadsto}_{\mathcal{R}} (S',E')$ for some E', and

2. S' is acceptable (wrt. $\mathcal{C}$).

■

It is worth mentioning that neither existence nor uniqueness of a successor state is guaranteed in general. Regarding uniqueness, the application of a fixed set of causal relationships is known to be order independent.[8] Yet a different ordering may allow the application of a different set of relationships, in which case the resulting successor states need not coincide. This characterizes actions that are non-deterministic as regards their indirect effects. If no successor state at all exists although one or more action laws are applicable, then this indicates that the action in question has additional, *implicit* preconditions [Ginsberg and Smith, 1988b, Lin and Reiter, 1994] which are not met.

Example 1 (continued) Performing $\mathtt{toggle(s_1)}$ in state S_0 (c.f. equation (9)) results in the intermediate state-effect pair

$$(S_1, \{\mathtt{closed(s_1)}\})$$

(where S_1 is as in equation (11)) according to the action laws for $\mathtt{toggle}(x)$ (c.f. equation (10)). The only applicable chain of causal relationships which results in a state that satisfies all underlying constraints, is the following:

$$\begin{array}{rlll}
\mathtt{closed(s_1)} & \underline{\text{causes}} & \mathtt{active(re_1)} & \underline{\text{if}} \;\; \neg\mathtt{ab(r_1)} \wedge \neg\mathtt{ab(re_1)} \\
\mathtt{active(re_1)} & \underline{\text{causes}} & \mathtt{closed(s_2)} & \\
\mathtt{closed(s_2)} & \underline{\text{causes}} & \mathtt{active(re_2)} & \underline{\text{if}} \;\; \neg\mathtt{ab(r_2)} \wedge \neg\mathtt{ab(re_2)} \\
\mathtt{active(re_2)} & \underline{\text{causes}} & \mathtt{closed(s_3)} & \\
\mathtt{closed(s_3)} & \underline{\text{causes}} & \mathtt{active(li)} & \underline{\text{if}} \;\; \neg\mathtt{ab(li)}
\end{array}$$

The successor state thus obtained is

$$\begin{array}{rl}
S' = \{ & \mathtt{closed(s_1)}, \neg\mathtt{closed(s_1')}, \mathtt{closed(s_2)}, \ldots, \\
& \mathtt{active(li)}, \neg\mathtt{active(li_1)}, \ldots, \\
& \mathtt{active(re_1)}, \mathtt{active(re_2)}, \\
& \neg\mathtt{ab(li)}, \ldots, \neg\mathtt{ab(r_1)}, \neg\mathtt{ab(r_2)}, \neg\mathtt{ab(re_1)}, \neg\mathtt{ab(re_2)} \,\}
\end{array} \tag{12}$$

■

Causal relationships help addressing the two key issues of the Ramification Problem. The commonsense law of persistence is weakened by further manipulating the state resulting from the application of this law and, by virtue of being directed relations, causal relationships allow to tell apart causal from mere logical consequences of state constraints. For a more detailed discussion on these and other aspects of the theory of causal relationships, including a thorough comparison with related approaches to the Ramification Problem, we refer the reader to [Thielscher, 1997b].

[8] Proposition 7 in [Thielscher, 1997b]

4 Dynamic Diagnosis

The framework introduced in the previous section provides means to give formal specifications of dynamic systems. The *static* part of such a specification fixes the entities and fluent names, and it also describes the static relations among the fluents in form of state constraints. The *dynamic* part specifies the actions which can be used to manipulate the system's state. These definitions are accompanied by action laws, focusing on the direct effects, and by causal relationships, concerning the indirect effects. Our theory of dynamic diagnosis to be developed next builds on this framework.

To begin with, descriptions of dynamic systems which are subject to diagnosis are assumed to include and employ the special fluent name `ab` to represent any aspect of abnormality in the system. The intuition is that usually all instances of `ab` are false. If, however, the system does not exhibit its regular behavior, then this can be accounted for by one or more affirmative instances of `ab`.

Our theory of dynamic diagnosis respects available knowledge of the *a priori* likelihood of component failures. Since it is often difficult if not impossible to provide precise numerical knowledge of probabilities, the theory accepts qualitative rather than quantitative information. Moreover, it does not rely on complete knowledge. Possibly incomplete information as to different degrees of abnormality is formally represented by a partial ordering, denoted $<$, among the instances of fluent `ab`.[9] If, for instance, we specify that $\mathtt{ab}(\mathtt{li}) < \mathtt{ab}(\mathtt{r_1})$, then this indicates that a broken light bulb `li` is *a priori* more likely than resistor $\mathtt{r_1}$ being out of order. Being a partial ordering, the comparison relation $<$ may be indifferent regarding certain instances of `ab`. The extreme is the empty relation, in which case diagnosing must be performed from very first principles. Thus our theory assumes that all abnormalities have equal *a priori* likelihood unless explicitly stated otherwise.[10]

Definition 6 A *system description* is a tuple $(\mathcal{E}, \mathcal{F}, \mathcal{A}, \mathcal{L}, \mathcal{C}, \mathcal{R}, <)$ consisting of entities, fluent and action names, action laws, state constraints, causal relationships, and a partial ordering on the set of ground instances of $\mathtt{ab} \in \mathcal{F}$. ■

Example 1 (continued) Our example system can be described by the 7-tuple SD_1 consisting of

- the 15 entities and 3 fluent names as introduced in the preceding section;

[9] Partial orderings are binary relations which are irreflexive, antisymmetric, and transitive. These orderings are *strict* if they relate any pair of disjoint elements either way. Later in this paper we refer to the notion of strict orderings (written $\ll$) *extending* a partial one (say, $<$), which means that $a \ll b$ whenever $a < b$.

[10] In particular, we do not try to deduce qualitative knowledge of *a priori* likelihood from state constraints, just because these are known to provide insufficient causal information [Pearl, 1988].

- the unary action name `toggle` accompanied by the two action laws of equation (10);
- the state constraints and causal relationships of Figure 2 and 3, respectively;
- the knowledge that both light bulbs and relays are more likely to break than resistors, i.e., the following partial ordering:

$$\mathtt{ab}(c_1) < \mathtt{ab}(c_2)$$
$$\text{for each } (c_1, c_2) \in \{\mathtt{li}, \mathtt{li}_1, \mathtt{li}_2, \mathtt{li}_3, \mathtt{re}_1, \mathtt{re}_2\} \times \{\mathtt{r}_1, \mathtt{r}_2, \mathtt{r}_3\}$$

■

System descriptions are used to specify the general static and dynamic properties of systems. These description form the basis for diagnosis problems, which are particular scenarios in which certain observations suggest that the system does not exhibit its regular behavior. Observations in classical diagnosis concern a unique state of the system. Usually they describe the state of the system only partially, in particular as far as abnormalities are concerned. Diagnosis then amounts to completing these partial descriptions, if possible. In dynamic diagnosis, observations may refer to system states at different stages, that is, prior, during, or after the performance of sequences of actions. The diagnosis task then is to draw the right conclusions from these situation-dependent observations, and in particular to propose diagnoses in case the observations suggest some abnormal behavior. Formally, observations are fluent formulas attached to a particular action sequence after whose performance the formula has been observed true.

Definition 7 Let *SD* be a system description. An *observation* is an expression

$$F \ \underline{\text{after}} \ [a_1, \ldots, a_n]$$

where F is a closed fluent formula and each of $a_1, \ldots, a_n$ is an action ($n \geq 0$). A *diagnosis problem* is a pair $(SD, \mathcal{O})$ consisting of a system description SD and a set of observations $\mathcal{O}$. ■

Example 1 (continued) The observation

$$\neg\mathtt{closed}(\mathtt{s}_1) \wedge \neg\mathtt{closed}(\mathtt{s}'_1) \wedge \neg\mathtt{closed}(\mathtt{s}'_2) \wedge \neg\mathtt{closed}(\mathtt{s}'_3) \quad \underline{\text{after}} \ [\,] \tag{13}$$

constitutes a partial description of the initial state of our circuit as depicted in Figure 1. Suppose it has further been observed that light bulb `li` stays off after toggling switch $\mathtt{s}_1$. This we can formally express as

$$\neg\mathtt{active}(\mathtt{li}) \ \underline{\text{after}} \ [\mathtt{toggle}(\mathtt{s}_1)] \tag{14}$$

■

It has been said that diagnosing amounts to drawing the right conclusions from the given observations and on the basis of the formal system description. We are now prepared for a precise definition of this task. In general, the observations that constitute a diagnosis problem provide only incomplete information as to the entire state of affairs. This is especially true if non-deterministic actions are involved, because then complete information means to know the actual result of any possible sequence of non-deterministic actions. One therefore has to expect that there be more than just one unique state of affairs that fits the observations. Following standard terminology in logic, we call any possible state of affairs an *interpretation*, and if the latter accounts for all given observations, then it is called a *model*.

Interpretations are constructed on the basis of a branching time structure, where each branch represents the performance of a particular action sequence and is rooted in the initial state of the system. An interpretation therefore must not just tell us exactly what happens during the execution of one particular sequence of actions. Rather it needs to provide this information as to any possible course of events. This supports so-called hypothetical reasoning about actions, which in turn helps with finding further actions to be taken towards fully determining the cause of an observed system failure. Of course we assume the system always evolves according to the underlying action laws and causal relationships. That is to say, whenever some state S results from performing some action sequence, and some further action a is executed, then the result should be a successor of S and a.

Definition 8 Let $(SD, \mathcal{O})$ be a diagnosis problem. The *transition model* Σ of SD is a mapping from state-action pairs to (possibly empty) sets of states such that $\Sigma(S, a)$ is defined iff S is acceptable,[11] and $S' \in \Sigma(S, a)$ iff S' is a successor of S and a.

An *interpretation* for $(SD, \mathcal{O})$ is a pair (Res, Σ) where Σ is the transition model of SD and Res is a partial mapping from finite action sequences (including the empty one) to acceptable states such that

1. $Res([\,])$ is defined;
2. for any sequence $a^* = [a_1, \ldots, a_{k-1}, a_k]$ of actions ($k > 0$),
   (a) $Res(a^*)$ is defined iff so is $Res([a_1, \ldots, a_{k-1}])$ and the set $\Sigma(Res([a_1, \ldots, a_{k-1}]), a_k)$ is not empty, and
   (b) $Res(a^*) \in \Sigma(Res([a_1, \ldots, a_{k-1}]), a_k)$.

■

Example 1 (continued) Let SD_1 be as above, and let Σ_1 be its transition model as determined by the underlying action laws and

[11] Recall that acceptable states are those which satisfy all state constraints.

causal relationships. Our electric circuit is deterministic, that is, for each acceptable state S and each action a the set $\Sigma_1(S, a)$ of successor states is either empty or singleton. In deterministic systems interpretations are uniquely characterized by the initial state, $Res([\,])$. In this way, setting $Res([\,]) = S_0$ (c.f. equation (9)) determines one out of many possible interpretations for a diagnosis problem in our example system. ■

Interpretations always tell us the exact result of performing any executable action sequence.[12] It is therefore straightforward to determine whether an observation is true with regard to a particular interpretation. First of all, it can be true only if the state is defined which results from performing the sequence of actions in question. If, moreover, the fluent formula in question is true in that state, then the observation itself is true. This naturally leads to the definition of models, which are interpretations in which all observations made in a diagnosis problem are true.

Definition 9 Let (Res, Σ) be an interpretation for a diagnosis problem $(SD, \mathcal{O})$. An observation F after $[a_1, \ldots, a_n]$ $(n \geq 0)$ is *true* in this interpretation iff $Res([a_1, \ldots, a_n])$ is defined and F is true in $Res([a_1, \ldots, a_n])$. An interpretation I is a *model* for a diagnosis problem $(SD, \mathcal{O})$ iff all observations in $\mathcal{O}$ are true in I. ■

Example 1 (continued) Let SD_1 and Σ_1 be as above, and let the interpretation $I = (Res, \Sigma_1)$ be determined by $Res([\,]) = S_0$. Observation (13) is true in I since all observable switches are indeed open in the initial state S_0. Observation (14), on the other hand, is false in I. For the only successor of S_0 and $\mathtt{toggle(s_1)}$ is S' of equation (12), in which $\neg\mathtt{active(li)}$ is false. There exist a number of other interpretations in which both observations are true, hence which are models of the corresponding diagnosis problem. Among them are models which correspond to the five diagnoses $d_1, \ldots, d_5$ (c.f. equation (6)), but there are also models whose initial states include many more abnormalities. ■

So far our theory does not treat the instances of fluent name `ab` any special. Any interpretation that fits the observations constitutes a model, regardless of the amount of abnormality it presupposes. What still needs to be done is to suitably reflect the intention of using abnormality fluents, namely, to assume normal circumstances to the largest possible extent. Put in other words, among all models for a diagnosis problem we are especially interested in those which are somehow minimally abnormal.

We have argued in Section 2 that minimization should be conducted initially and only once in order to overcome the specific dif-

[12] An action sequence a^* is called executable wrt. interpretation (Res, Σ) iff $Res(a^*)$ is defined.

ficulties which the phenomenon of causality raises in dynamic diagnosis. Minimization is formally achieved by a model preference criterion. Basically, models are preferable which declare false initially more instances of `ab` than other models. This strategy needs to be refined if the underlying system description includes qualitative prior knowledge of the likelihood of abnormalities. In this case instances of `ab` which are more unlikely are to be preferably minimized.

Definition 10 Let $(SD, \mathcal{O})$ be a diagnosis problem with partial ordering $<$. If $M = (Res, \Sigma)$ is a model for $(SD, \mathcal{O})$, then M is *preferred* iff we can find a strict ordering $\ll$ extending $<$ such that the following holds: For each model $M' = (Res', \Sigma)$ for $(SD, \mathcal{O})$ and each fluent $\mathtt{ab}(c) \in Res([\,]) \setminus Res'([\,])$, there is some $\mathtt{ab}(c') \in Res'([\,]) \setminus Res([\,])$ such that $\mathtt{ab}(c) \ll \mathtt{ab}(c')$ ■

In words, a preferred model is obtained by first choosing a minimization strategy, that is, a strict ordering which respects the given partial one. With the ordering fixed all models are preferred whose evolution function Res satisfies the following: Suppose some abnormality $\mathtt{ab}(c)$ is initially true in Res but false in the evolution function Res' of some other model. Then there must be another abnormality $\mathtt{ab}(c')$ which is of higher priority than $\mathtt{ab}(c)$ according to the chosen strict ordering and which is initially false in Res but true in Res'. Notice that the minimization strategy, i.e., the strict ordering, need not be unique, namely, in case the underlying partial ordering is truly partial. Different minimization strategies may lead to different preferred models, which all have to be considered equal thanks to the lack of more precise knowledge.

Example 1 (continued) Let SD_1 and Σ_1 be as above, and let $\mathcal{O}_1$ consist of the two observations (13) and (14). Each preferred model (Res, Σ_1) of $(SD_1, \mathcal{O}_1)$ satisfies exactly one of the following conditions:

1. $\mathtt{ab}(\mathtt{re}_1) \in Res([\,])$, and $\neg\mathtt{ab}(c) \in Res([\,])$ for all $c \neq \mathtt{re}_1$;
2. $\mathtt{ab}(\mathtt{re}_2) \in Res([\,])$, and $\neg\mathtt{ab}(c) \in Res([\,])$ for all $c \neq \mathtt{re}_2$;
3. $\mathtt{ab}(\mathtt{li}) \in Res([\,])$, and $\neg\mathtt{ab}(c) \in Res([\,])$ for all $c \neq \mathtt{li}$

Models that do not obey either of these conditions do not admit a strict ordering $\ll$ satisfying the requirements of Definition 10. To see why, let, for instance, $M_1 = (Res_1, \Sigma_1)$ denote a model where $\mathtt{ab}(\mathtt{r}_1) \in Res_1([\,])$. In order for M_1 to be preferred, each model that declares initially false $\mathtt{ab}(\mathtt{r}_1)$ should admit another abnormality instead. Moreover, this 'compensating' abnormality needs to be less preferred according to some self-chosen strict ordering—which, of course, must respect the given partial one. Now, there are models which declare initially false $\mathtt{ab}(\mathtt{r}_1)$. These models indeed each admit another abnormality, e.g., the ones whose initial states have $\mathtt{ab}(\mathtt{re}_1)$

as the only affirmative instance of $\mathtt{ab}$. But any strict ordering with $\mathtt{ab(r_1)} \ll \mathtt{ab(re_1)}$ violates the given $\mathtt{ab(re_1)} < \mathtt{ab(r_1)}$, which is why M_1 cannot be preferred.[13] ∎

Preferred models for a diagnosis problem provide what we are looking for, namely, the diagnoses. More specifically speaking, we can take as diagnosis any distribution of initial affirmative instances of fluent ab if this distribution occurs in at least one preferred model. Following standard terminology, the notion of preferred model also allows a more general definition of what conclusions can be drawn from a formal diagnosis problem.

Definition 11 Let $(SD, \mathcal{O})$ be a diagnosis problem. An observation is *entailed* by $(SD, \mathcal{O})$ iff it is true in all preferred models for $(SD, \mathcal{O})$. ∎

This entailment relation is obviously truly nonmonotonic in that adding observation to a diagnosis problem may disable previously valid entailments. It thus does not enjoy the property of restricted monotonicity of [Lifschitz, 1993]. This property is indeed undesired once state descriptions include fluents representing abnormalities [Thielscher, 1996a].

The following result shows that our theory of entailment solves the problem elaborated in Sectioon 2.

Theorem 12 *Let SD_1 be the system description of the circuit of Figure 1 as used throughout this section, and let $\mathcal{O}$ consist of the observation*

$$\neg\mathtt{closed(s_1)} \wedge \neg\mathtt{closed(s'_1)} \wedge \neg\mathtt{closed(s'_2)} \wedge \neg\mathtt{closed(s'_3)} \wedge \mathtt{ab(r_2)} \ \underline{\text{after}}\ [\,]$$

Then $(SD_1, \mathcal{O})$ entails $\mathtt{ab(re_2)}$ $\underline{\text{after}}$ $[\mathtt{toggle(s_1)}]$.

Proof: It is consistent with the observation to assume that the given $\mathtt{ab(r_2)}$ is the only initial abnormality. All preferred models (Σ, Res) therefore coincide as far as abnormality in $Res([\,])$ is concerned. In particular, we know that $\neg\mathtt{ab(r_1)} \wedge \neg\mathtt{ab(re_1)}$ is true in each such $Res([\,])$. Therefore, according to the transition model of SD_1, after closing switch $\mathtt{s_1}$ the topmost causal relationship in Figure 3 applies and activates relay $\mathtt{re_1}$, which in turn causes $\mathtt{closed(s_2)}$ and, hence, $\mathtt{ab(re_2)}$ following the fifth causal relationship from the bottom and the bottommost, respectively, of Figure 3. Thus we know that $\mathtt{ab(re_2)} \in Res([\mathtt{toggle(s_1)}])$ holds in all preferred models. Hence, $\mathtt{ab(re_2)}$ $\underline{\text{after}}$ $[\mathtt{toggle(s_1)}]$ is entailed. ∎

[13] The reader should notice that we have obtained the above three diagnoses in the light of resistors being *a priori* more unlikely to break than relays or light bulbs. Had we had to diagnose from first principles, three more preferred models would have been obtained, each of which assesses one of $\mathtt{ab(r_i)}$ ($i = 1, 2, 3$).

Entailed observations need not be restricted to the sequence of actions that has actually been performed. In particular, they may refer to actions possibly taken in the future. In diagnosis problems, this kind of hypothetical reasoning may help with suggesting actions to be taken towards fully determining the cause for an abnormal behavior in case the given observations do not entail a unique conjunction of affirmative instances of $\mathtt{ab}$. The diagnosis problem may additionally entail observations that indicate under what circumstances a more definite conclusion would be possible. We conclude this section with a formalization of the diagnosis process described in the introduction, where the active production of observations helped the diagnostician come to a suitable unique diagnosis.

Example 2 Let SD_1 be as above, and let $\mathcal{O}_2$ consist solely of observation (13). Then it is consistent to assume away all abnormalities. Consequently, the observation

$$\mathtt{active(li)} \ \underline{\text{after}} \ [\mathtt{toggle(s_1)}]$$

is true in all preferred models, hence is entailed. This conclusion is invalidated if the observation $\neg\mathtt{active(li)}$ $\underline{\text{after}}$ $[\mathtt{toggle(s_1)}]$ is added. The modified diagnosis problem entails

$$\mathtt{ab(re_1)} \vee \mathtt{ab(re_2)} \vee \mathtt{ab(li)} \ \underline{\text{after}} \ [\,]$$

which indicates the three possible diagnoses for this problem. (In fact, a stronger conclusion is entailed, namely, that these three diagnoses are pairwise exclusive.)

According to the underlying transition model, the diagnosis problem $(SD_1, \mathcal{O}_2)$ also entails the observation

$$\begin{array}{c} \mathtt{active(li_3)} \supset \mathtt{ab(li)} \wedge \neg\mathtt{ab(re_1)} \wedge \neg\mathtt{ab(re_2)} \\ \underline{\text{after}} \\ [\mathtt{toggle(s_1)}, \mathtt{toggle(s'_3)}] \end{array} \tag{15}$$

That is to say, if toggling switch $\mathtt{s'_3}$ activated light $\mathtt{li_3}$, then the diagnosis problem would admit a unique solution, namely, the diagnosis that bulb $\mathtt{li}$ is broken. (To see why observation (15) is entailed, notice first that light bulb $\mathtt{li_3}$ can only be on if switch $\mathtt{s_3}$ is closed. But then light $\mathtt{li}$ has been observed off after toggling $\mathtt{s_1}$, which is possible just in case $\mathtt{ab(li)}$ held initially. The three original diagnoses being exclusive, it follows that the two relays $\mathtt{re_1}$ and $\mathtt{re_2}$, respectively, are—by default—in order.) ■

5 A Calculus

In this section, we briefly describe a suitable action calculus which is capable of handling exceptions to ramification in precisely the way our theory of dynamic diagnosis suggests how it should be done.

Our encoding of diagnosis problems builds on results described in preceding papers on both the Ramification and Qualification Problem [Thielscher, 1997b, Thielscher, 1996a, Thielscher, 1996b]. These axiomatizations all employ the representation technique underlying the *Fluent Calculus* [Hölldobler and Schneeberger, 1990, Thielscher, 1997b]. As opposed to the Situation Calculus [McCarthy and Hayes, 1969], the former employs structured state terms which each consists in a collection of all fluent literals that are true in the state being represented. To this end, fluent literals are reified, i.e., formally represented as terms. An example state term is $\neg\mathtt{closed}(\mathtt{s_1}) \circ \mathtt{active}(\mathtt{li}) \circ \neg\mathtt{ab}(\mathtt{re_1}) \circ \ldots$ where the negation sign denotes a special unary function and $\circ$ a special binary function which obeys the laws of associativity and commutativity. It has first been argued in [Hölldobler and Schneeberger, 1990] that this representation technique, which appeals exclusively to classical, i.e., monotonic logic, avoids extra axioms to encode the general commonsense law of persistence: The effects of actions are modeled by manipulating state terms through removal and addition of sub-terms. Then all sub-terms which are not affected by these operations remain in the state term, hence continue to be true. In this way the Fluent Calculus provides a uniform solution to both the representational and the inferential aspect of the Frame Problem.

In [Thielscher, 1997b], we have presented a Fluent Calculus-based axiomatization of the theory of causal relationships. This axiomatization has been proved correct, as has the axiomatization described in [Thielscher, 1996a, Thielscher, 1996b], which embeds the former in Default Logic [Reiter, 1980] to address the Qualification Problem. The use of Default Logic can be straightforwardly adapted to the theory proposed in the present paper.[14] To this end, this open default rule (which represents all of its ground instances) is introduced:

$$\frac{:\ \forall s\, [\, \mathit{Initial}(s) \supset \neg \mathit{holds}(\mathtt{ab}(x), s)\,]}{\forall s\, [\, \mathit{Initial}(s) \supset \neg \mathit{holds}(\mathtt{ab}(x), s)\,]}$$

It should be read as follows: Provided it is consistent, conclude that if s encodes the initial state then an instance $\mathtt{ab}(c)$ is false in s. Additionally, to minimize certain abnormalities with higher priority if necessary, we employ the concept of *Prioritized Default Logic* [Brewka, 1994]. The resulting axiomatization is provably correct wrt. the formal theory developed in the preceding section. That is to say, in the corresponding default theory the encoding of an observation is skeptically entailed (see [Reiter, 1980]) if and only if the abstract diagnosis problem entails the observation according to Definition 11. This correctness result follows directly from the results and proofs in [Thielscher, 1996b], to which we refer the reader for full details.

[14] It should be stressed that the Fluent Calculus provides monotonic solutions to both the Frame Problem as well as the Ramification Problem. Yet both the Qualification Problem and, as we have seen, the problem of ramifications having exceptions necessitate some kind of nonmonotonicity.

6 Discussion

We have proposed a formal theory of dynamic diagnosis, where the system under examination is subject to actions, e.g. performed by the diagnostician. This dynamic aspect we have captured by appealing to Action Theory. The behavior of a system under healthy condition is specified by means of state constraints. These formulas, static by nature, give rise to indirect effects once the dynamic aspect enters. Diagnosis is required in case the system does not exhibit its regular behavior. In terms of Action Theory, this amounts to accounting for exceptions to both state constraints and the ramifications they trigger. We have illustrated that the dynamic aspect raises a new challenge for formal theories of diagnosis, which is due to the phenomenon of causality. We have shown how this challenge can be met on the basis of the theory of causal relationships. To this end, we have taken abnormalities as fluents which are assumed false initially but may be indirectly affected by the execution of actions. Besides exploiting causal information when generating the most plausible diagnoses, our theory also takes into account possibly available, qualitative knowledge of the *a priori* likelihood of components to fail. For the entire theory there exists a provably correct axiomatization based on the Fluent Calculus paradigm and which uses Default Logic to accommodate the nonmonotonic aspect of the diagnostic problem.

We have chosen the term “dynamic” solely to reflect the fact that the systems under investigation may exhibit different states in the course of time, as a consequence of actions. While in the diagnosis community the notion of “dynamic diagnosis” usually refers to the analysis of self-evolving systems, recent work in Action Theory (e.g., [Thielscher, 1995, Reiter, 1996, Shanahan, 1997], just to mention a few) showed that the gap is less deep than one might expect between dynamic systems which idle unless actions are performed, and those that are self-evolving. In particular, the theory of causal relationships, along with its axiomatization on the basis of the Fluent Calculus, has recently been extended to allow for natural events aside from exogenous, volitional actions [Thielscher, 1998]. By nature causal relationships apply whenever some effect occurs, no matter whether the latter is a consequence of an exogenous action or of an internal event.

The main concern of [Thielscher, 1998] is the problem of minimizing ecent occurrences wrt. formal scenario descriptions. Coupling this work with the result of the present paper would yield a generalized theory of so-called *event-based* dynamic diagnosis, where part of a diagnosis is a sequence of events according to which the system supposedly has evolved. This would bring our work closer to that of [Cordier and Thiébaux, 1994]. The main conceptual difference is that the latter work is mostly defined on explicit state transition models while we started off from compact and concise specifications of dynamic systems in terms of effect descriptions, state constraints, and causal relationships. The authors of [Cordier and Thiébaux, 1994] themselves stress the importance of dealing with such compact specifications and in particular of a satisfactory solution to the Ramification Problem, which the theory of causal relationships provides [Thielscher, 1997b].

Action Theory as the basis for dynamic diagnosis has been independently proposed in [McIlraith, 1997a, McIlraith, 1997b]. There a Situation Calculus-based axiomatization of actions and their direct and indirect effects is used directly to formalize and solve dynamic diagnosis problems. In comparison to our theory, a restriction is imposed on the diagnosis problems which can be expressed due to the restricted form of state constraints the theory supports, namely, which need to form a so-called *stratified* theory. A second restriction stems from the fact that diagnosis is performed from first principles; knowledge as to the *a priori* likelihood of particular abnormalities is not supported. Minimization in this approach is used for two different purposes; first, to assume away abnormal exceptions to state constraints and ramifications and, second, to solve the Ramification Problem itself. Care has therefore to be taken that these two minimization steps do not interfere. As a solution, the minimization accounting for indirect effects is performed in a pre-processing step. Thus the computation of indirect effects is 'compiled' into the action laws. In this way the Ramification Problem is circumvented for the price of a potentially redundant axiomatization. Arguments in favor of the theory of causal relationships as a solution to the Ramification Problem and a thorough comparison to other approaches can be found in [Thielscher, 1997b].

Generally, the problem of ramifications having exceptions has received little attention in literature up to now, probably because satisfactory solutions to the Ramification Problem itself have not emerged until very recently. To the best of the author's knowledge, the only existing papers dealing with qualifications of ramifications are [Baral and Lobo, 1997, Zhang, 1996]. In both of them expressions resembling causal relationships are allowed to be defeasible. The first of the approaches suffers from a rather formal, hence less intuitive definition of successor states, which essentially relies on the theory of answer sets in extended logic programs [Gelfond and Lifschitz, 1991]. This makes it difficult to verify the author's claim that their approach does respect causal information when minimizing abnormality. On the other hand, the authors admit that their approach, as it stands, is restricted to deterministic system descriptions, as opposed to our's. The second of the aforementioned approaches, [Zhang, 1996], does not go beyond defining a notion of successor state based on minimizing abnormality. If this approach shall be applied to dynamic diagnosis, then measures need to be taken in order not to getting caught in the causality trap illustrated with our key example of Section 2.

Acknowledgments

The author is grateful to Marie-Odile Cordier for her comments and pointers to relevant literature, which motivated the revision of an earlier version of the paper in several important respects. The author also wants to thank Rob Miller and Wolfgang Nejdl for helpful questions and remarks.

References

[Baral and Lobo, 1997] Chitta Baral and Jorge Lobo. Defeasible specifications in action theories. In M. E. Pollack, editor, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1441–1446, Nagoya, Japan, August 1997. Morgan Kaufmann.

[Brewka, 1994] Gerhard Brewka. Adding priorities and specificity to default logic. In C. MacNish, D. Pearce, and L. M. Pereira, editors, *Proceedings of the European Workshop on Logics in AI (JELIA)*, volume 838 of *LNAI*, pages 50–65. Springer, September 1994.

[Cordier and Thiébaux, 1994] Marie-Odile Cordier and Sylvie Thiébaux. Event-based diagnosis for evolutive systems. In *Proceedings of the International Workshop on Principles of Diagnosis (DX)*, New Palz, NY, 1994. (Also available as IRISA Internal Report 819, Rennes Cedex, France).

[Elkan, 1992] Charles Elkan. Reasoning about action in first-order logic. In *Proceedings of the Conference of the Canadian Society for Computational Studies of Intelligence (CSCSI)*, pages 221–227, Vancouver, Canada, May 1992. Morgan Kaufmann.

[Gelfond and Lifschitz, 1991] Michael Gelfond and Vladimir Lifschitz. Classical Negation in Logic Programs and Disjunctive Databases. *New Generation Computing*, 9:365–385, 1991.

[Ginsberg and Smith, 1988a] Matthew L. Ginsberg and David E. Smith. Reasoning about action I: A possible worlds approach. *Artificial Intelligence*, 35:165–195, 1988.

[Ginsberg and Smith, 1988b] Matthew L. Ginsberg and David E. Smith. Reasoning about action II: The qualification problem. *Artificial Intelligence*, 35:311–342, 1988.

[Hanks and McDermott, 1987] Steve Hanks and Drew McDermott. Nonmonotonic logic and temporal projection. *Artificial Intelligence*, 33(3):379–412, 1987.

[Hölldobler and Schneeberger, 1990] Steffen Hölldobler and Josef Schneeberger. A new deductive approach to planning. *New Generation Computing*, 8:225–244, 1990.

[Lifschitz, 1987] Vladimir Lifschitz. Formal theories of action (preliminary report). In J. McDermott, editor, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 966–972, Milan, Italy, August 1987. Morgan Kaufmann.

[Lifschitz, 1990] Vladimir Lifschitz. Frames in the space of situations. *Artificial Intelligence*, 46:365–376, 1990.

[Lifschitz, 1993] Vladimir Lifschitz. Restricted monotonicity. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, pages 432–437, Washington, DC, July 1993.

[Lin and Reiter, 1994] Fangzhen Lin and Ray Reiter. State constraints revisited. *Journal of Logic and Computation*, 4(5):655–678, 1994.

[Lin, 1995] Fangzhen Lin. Embracing causality in specifying the indirect effects of actions. In C. S. Mellish, editor, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1985–1991, Montreal, Canada, August 1995. Morgan Kaufmann.

[McCain and Turner, 1995] Norman McCain and Hudson Turner. A causal theory of ramifications and qalifications. In C. S. Mellish, editor, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1978–1984, Montreal, Canada, August 1995. Morgan Kaufmann.

[McCarthy and Hayes, 1969] John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4:463–502, 1969.

[McCarthy, 1977] John McCarthy. Epistemological problems of artificial intelligence. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1038–1044, Cambridge, MA, 1977. MIT Press.

[McIlraith, 1997a] Sheila A. McIlraith. Representing actions and state constraints in model-based diagnosis. In B. Kuipers and B. Webber, editors, *Proceedings of the AAAI National Conference on Artificial Intelligence*, pages 43–49, Providence, RI, July 1997. MIT Press.

[McIlraith, 1997b] Sheila A. McIlraith. *Towards a Formal Account of Diagnostic Problem Solving.* PhD thesis, Department of Computer Science, University of Toronto, 1997.

[Pearl, 1988] Judea Pearl. Embracing causality in default reasoning. *Artificial Intelligence*, 35(2):259–271, 1988.

[Reiter, 1980] Ray Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

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

[Reiter, 1996] Ray Reiter. Natural actions, concurrency and continuous time in the situation calculus. In L. C. Aiello, J. Doyle, and S. Shapiro, editors, *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 2–13, Cambridge, MA, November 1996. Morgan Kaufmann.

[Shanahan, 1995] Murray Shanahan. Default reasoning about spatial occupancy. *Artificial Intelligence*, 74:147–163, 1995.

[Shanahan, 1997] Murray Shanahan, editor. *Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia*. MIT Press, 1997.

[Thielscher, 1995] Michael Thielscher. The logic of dynamic systems. In C. S. Mellish, editor, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1956–1962, Montreal, Canada, August 1995. Morgan Kaufmann.

[Thielscher, 1996a] Michael Thielscher. Causality and the qualification problem. In L. C. Aiello, J. Doyle, and S. Shapiro, editors, *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 51–62, Cambridge, MA, November 1996. Morgan Kaufmann.

[Thielscher, 1996b] Michael Thielscher. Qualification and Causality. Technical Report TR-96-026, International Computer Science Institute (ICSI), Berkeley, CA, July 1996. (Electronically available. Submitted for publication).

[Thielscher, 1997a] Michael Thielscher. Qualified ramifications. In B. Kuipers and B. Webber, editors, *Proceedings of the AAAI National Conference on Artificial Intelligence*, pages 466–471, Providence, RI, July 1997. MIT Press.

[Thielscher, 1997b] Michael Thielscher. Ramification and causality. *Artificial Intelligence*, 89(1–2):317–364, 1997.

[Thielscher, 1998] Michael Thielscher. How (not) to minimize events. In A. G. Cohn and L. K. Schubert, editors, *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR)*, Trento, Italy, June 1998. Morgan Kaufmann.

[Zhang, 1996] Yan Zhang. Compiling causality into action theories. In *Proceedings of the Symposium on Logical Formalizations of Commonsense Reasoning*, pages 263–270, January 1996.