Diagnosis of plan execution and the executing agent

Nico Roos[1] and Cees Witteveen[2,3]

[1] Dept of Computer Science, Universiteit Maastricht
P.O.Box 616, NL-6200 MD Maastricht
roos@cs.unimaas.nl

[2] Faculty EEMCS, Delft University of Technology
P.O.Box 5031, NL-2600 GA Delft
witt@ewi.tudelft.nl

[3] Centre of Mathematics and Computer Science
P.O. Box 94079, NL-1090 GB Amsterdam
C.Witteveen@cwi.tudelft.nl

Abstract. We adapt the Model-Based Diagnosis framework to perform (agent-based) plan diagnosis. In plan diagnosis, the system to be diagnosed is a plan, consisting of a partially ordered set of instances of actions, together with its executing agent. The execution of a plan can be monitored by making partial observations of the results of actions. Like in standard model-based diagnosis, observed deviations from the expected outcomes are explained qualifying some action instances that occur in the plan as behaving abnormally. Unlike in standard model-based diagnosis, however, in plan diagnosis we cannot assume that actions fail independently. We focus on two sources of dependencies between failures: dependencies that arise as a result of a malfunction of the executing agent, and dependencies that arise because of dependencies between action instances occurring in a plan. Therefore, we introduce causal rules that relate health states of the agent and health states of actions to abnormalities of other action instances. These rules enable us to introduce causal set and causal effect diagnoses that use the underlying causes of plan failing to explain deviations and to predict future anomalies in the execution of actions.

1 Introduction

In Model-Based Diagnosis (MBD) a model of a system consisting of components, their interrelations and their behavior is used to establish why the system is malfunctioning. Plans resemble such system specifications in the sense that plans also consist of components (action specifications), their interrelations and a specification of the (correct) behavior of each action. Based on this analogy, the aim of this paper is to adapt and extend a classical Model-Based Diagnosis (MBD) approach to the diagnosis of plans.

To this end, we will first formally model a plan consisting of a partially ordered set of actions as a system to be diagnosed, and subsequently we will describe how a diagnosis can be established using *partial observations* of a plan in progress. Distinguishing between normal and abnormal execution of actions in a plan, we will introduce sets of

actions qualified as abnormal to explain the deviations between expected plan states and observed plan states. Hence, in this approach, a plan diagnosis is just a set of abnormal actions that is able to explain the deviations observed. Although plan diagnosis conceived in this way is a rather straightforward application of MBD to plans, we do need to introduce new criteria for selecting acceptable plan diagnoses: First of all, while in standard MBD usually subset-minimal diagnoses, or within them *minimum (cardinality)* diagnoses, are preferred, we also prefer *maximum informative* diagnoses. The latter type of diagnosis maximizes the exact similarity between predicted and observed plan states. Although maximum informative diagnoses are always subset minimal, they are not necessarily of minimum cardinality. More differences between MBD and plan diagnosis appear if we take a more detailed look into the reasons for choosing minimal diagnoses. The idea of establishing a minimal diagnosis in MBD is governed by the principle of *minimal change*: explain the abnormalities in the behavior observed by changing the qualification from normal to abnormal for as few system components as necessary. Using this principle is intuitively acceptable if the components qualified as abnormal are failing *independently*. However, as soon as *dependencies* exist between such components, the choice for minimal diagnoses cannot be justified. As we will argue, the existence of dependencies between failing actions in a plan is often the rule instead of an exception. Therefore, we will refine the concept of a plan diagnosis by introducing the concept of a *causal diagnosis*. To establish such a causal diagnosis, we consider both the executing agent and its plan as constituting the system to be diagnosed and we explicitly relate health states of the executing agent and subsets of (abnormally qualified) actions to the abnormality of other actions in the form of causal rules. These rules enable us to replace a set of dependent failing actions (e.g. a plan diagnosis) by a set of unrelated *causes* of the original diagnosis. This independent and usually smaller set of causes constitutes a causal diagnosis, consisting of a health state of an agent and an independent (possibly empty) set of failing actions. Such a causal diagnosis always generates a cover of a minimal diagnosis. More importantly, such causal diagnoses can also be used to predict failings of actions that have to be executed in the plan and thereby also can be used to assess the consequences of such failures for goal realizability.

This paper is organized as follows. Section 2 introduces the preliminaries of plan-based diagnosis, while Section 3 formalizes plan-based diagnosis. Section 4 extends the formalization to determining the agent's health state. Finally, we briefly discuss some computational aspects of (causal) plan diagnosis. In Section 6, we place our approach into perspective by discussing some related approaches to plan diagnosis. and Section 7 concludes the paper.

2 Preliminaries

Model based Diagnosis In Model-Based Diagnosis (MBD) [4, 5, 12] a system S is modeled as consisting of a set $Comp$ of components and their relations, for each component $c \in Comp$ a set H_c of *health modes* is distinguished and for each health mode $h_c \in H_c$ of each component c a specific (input-output) behavior of c is specified. Given some input to S, its output is defined if the health mode of each component $c \in Comp$ is known. The diagnostic engine is triggered whenever, under the assumption

that all components are functioning normally, there is a discrepancy between the output as predicted from the input observations, and the actually observed output. The result of MBD is a suitable assignment of health modes to the components, called a *diagnosis*, such that the actually observed output is *consistent* with this health mode qualification or can be *explained* by this qualification. Usually, in a diagnosis one requires the number of components qualified as abnormally to be minimized.

States and Partial States We consider plan-based diagnosis as a simple extension of the model-based diagnosis where the model is not a description of an underlying system but a *plan* of an agent. Before we discuss plans, we introduce a simplified state-based view on the world, assuming that for the planning problem at hand, the world can be simply described by a set $Var = \{v_1, v_2, \ldots, v_n\}$ of variables and their respective *value domains* D_i. A *state of the world* σ then is a value assignment $\sigma(v_i) = d_i \in D_i$ to the variables. We will denote a state simply by an element of $D_1 \times D_2 \times \ldots \times D_n$, i.e. an n-tuple of values. It will not always be possible to give a complete state description. Therefore, we introduce a *partial state* as an element $\pi \in D_{i_1} \times D_{i_2} \times \ldots \times D_{i_k}$, where $1 \leq k \leq n$ and $1 \leq i_1 < \ldots < i_k \leq n$. We use $Var(\pi)$ to denote the set of variables $\{v_{i_1}, v_{i_2}, \ldots, v_{i_k}\} \subseteq Var$ specified in such a state π. The value d_j of variable $v_j \in Var(\pi)$ in π will be denoted by $\pi(j)$. The value of a variable $v_j \in Var$ not occurring in a partial state π is said to be *unknown* (or unpredictable) in π, denoted by $\perp$. Including $\perp$ in every value domain D_i allows us to consider every partial state π as an element of $D_1 \times D_2 \times \ldots \times D_n$.

Partial states can be ordered with respect to their information content: Given values d and d', we say that $d \leq d'$ holds iff $d = \perp$ or $d = d'$. The containment relation $\sqsubseteq$ between partial states is the point-wise extension of $\leq$: π is said to be contained in π', denoted by $\pi \sqsubseteq \pi'$, iff $\forall j[\pi(j) \leq \pi(j')]$. Given a subset of variables $V \subseteq Var$, two partial states π, π' are said to be V*-equivalent*, denoted by $\pi =_V \pi'$, if for every $v_j \in V$, $\pi(j) = \pi'(j)$. We define the partial state π restricted to a given set V, denoted by $\pi \restriction V$, as the state $\pi' \sqsubseteq \pi$ such that $Var(\pi') = V \cap Var(\pi)$.

An important notion for diagnosis is the notion of *compatibility* between partial states. Intuitively, two states π and π' are said to be compatible if there is no essential disagreement about the values assigned to variables in the two states. That is, for every j either $\pi(j) = \pi'(j)$ or at least one of the values $\pi(j)$ and $\pi'(j)$ is undefined. So we define π and π' to be compatible, denoted by $\pi \approx \pi'$, iff $\forall j[\pi(j) \leq \pi'(j)$ *or* $\pi'(j) \leq \pi(j)]$. As an easy consequence we have, using the notion of V-equivalent states, $\pi \approx \pi'$ iff $\pi =_{Var(\pi) \cap Var(\pi')} \pi'$. Finally, if π and π' are compatible states they can be *merged* into the $\sqsubseteq$-least state $\pi \sqcup \pi'$ containing them both: $\pi \sqcup \pi'(j) = max_{\leq}\{\pi(j), \pi'(j)\}$.

Goals An (elementary) goal g of an agent specifies a set of partial states an agent wants to bring about using a plan. Here, we specify each such a goal g as a constraint, that is a relation over some product $D_{i_1} \times \ldots \times D_{i_k}$ of domains.

We say that a goal g is satisfied by a partial state π, denoted by $\pi \models g$, if the relation g contains at least one tuple $(d_{i_1}, d_{i_2}, \ldots, d_{i_k})$ such that $(d_{i_1}, d_{i_2}, \ldots d_{i_k}) \sqsubseteq \pi$. We assume each agent to have a set G of such elementary goals $g \in G$. We use $\pi \models G$ to denote that all goals in G hold in π, i.e. for all $g \in G$, $\pi \models g$.

Actions and action schemes An *action scheme* or plan operator α is represented as a function that replaces the values of a subset $V_\alpha \subseteq Var$ by other values, dependent

upon the values of another set $V'_\alpha \supseteq V_\alpha$ of variables. Hence, every action scheme α can be modeled as a (partial) function $f_\alpha : D_{i_1} \times \ldots \times D_{i_k} \rightarrow D_{j_1} \times \ldots \times D_{j_l}$, where $1 \leq i_1 < \ldots < i_k \leq n$ and $\{j_1, \ldots, j_l\} \subseteq \{i_1, \ldots, i_k\}$. The variables whose value domains occur in $dom(f_\alpha)$ will be denoted by $dom_{Var}(\alpha) = \{v_{i_1}, \ldots, v_{i_k}\}$ and, likewise $ran_{Var}(\alpha) = \{v_{j_1}, \ldots, v_{j_l}\}$. Note that it is required that $ran_{Var}(\alpha) \subseteq dom_{Var}(\alpha)$. This functional specification f_α constitutes the *normal* behavior of the action scheme, denoted by f_α^{nor}.

Example 1. Figure 1 depicts two states σ_0 and σ_1 (the white boxes) each characterized by the values of four variables v_1, v_2, v_3 and v_4. The partial states π_0 and π_1 (the gray boxes) characterize a subset of values in a (complete) state. Action schemes are used to model state changes. The domain of the action scheme α is the subset $\{v_1, v_2\}$, which are denoted by the arrows pointing to α. The range of α is the subset $\{v_1\}$, which is denoted by the arrow pointing from α. Finally, the dashed arrow denotes that the value of variable v_2 is not changed by operator(s) causing the state change. ■

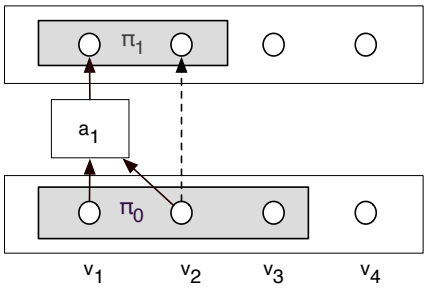

Fig. 1. Plan operators & states.

The correct execution of an action may fail either because of an inherent malfunctioning or because of a malfunctioning of an agent responsible for executing the action, or because of unknown external circumstances. In all these cases we would like to model the effects of executing such failed actions. Therefore, we introduce a set of *health modes* M_α for each action scheme α. This set M_α contains at least the normal mode nor, the mode ab indicating the most general abnormal behavior, and possibly several other specific fault modes. The most general abnormal behavior of action α is specified by the function f_α^{ab}, where $f_\alpha^{ab}(d_{i_1}, d_{i_2}, \ldots, d_{i_k}) = (\bot, \bot, \ldots, \bot)$ for every partial state $(d_{i_1}, d_{i_2}, \ldots, d_{i_k}) \in dom(f_\alpha)$.[4] To keep the discussion simple, in the sequel we distinguish only the health modes nor and ab.

Given a set $\mathcal{A}$ of action schemes, we will need to consider a set $A \subseteq inst(\mathcal{A})$ of *instances* of actions in $\mathcal{A}$. Such instances will be denoted by small roman letters a_i. If $type(a_i) = \alpha \in \mathcal{A}$, such an instance a_i is said to be of *type* α. If the context permits we will use "actions" and "instances of actions" interchangeably.

[4] This definition implies that the behavior of abnormal actions is essentially unpredictable.

Plans A plan is a tuple $P = \langle \mathcal{A}, A, < \rangle$ where $A \subseteq Inst(\mathcal{A})$ is a set of instances of actions occurring in $\mathcal{A}$ and $(A, <)$ is a partial order. The partial order relation $<$ specifies a precedence relation between these instances: $a < a'$ implies that the instance a must finish before the instance a' may start. We will denote the *transitive reduction* of $<$ by $\ll$, i.e., $\ll$ is the smallest subrelation of $<$ such that the transitive closure $\ll^+$ of $\ll$ equals $<$.

We assume that if in a plan P two action instances a and a' are independent, in principle they may be executed concurrently. This means that the dependency relation $<$ at least should capture all resource dependencies that would prohibit concurrent execution of actions. Therefore, we assume $<$ to satisfy the following *concurrency requirement*:

If $ran_{Var}(a) \cap dom_{Var}(a') \neq \varnothing$ then $a < a'$ or $a' < a$.[5]

That is, for concurrent instances, domains and ranges do not overlap.

Example 2. Figure 2 gives an illustration of a plan. Arrows relate the variables an action uses as inputs and the variables it produces as its outputs to the action itself. In this plan, the dependency relation is specified as $a_1 \ll a_3$, $a_2 \ll a_4$, $a_4 \ll a_5$, $a_4 \ll a_6$ and $a_1 \ll a_5$. Note that the last dependency has to be included because a_5 changes the value of v_2 needed by a_1. The action a_1 shows that not every variable occurring in the

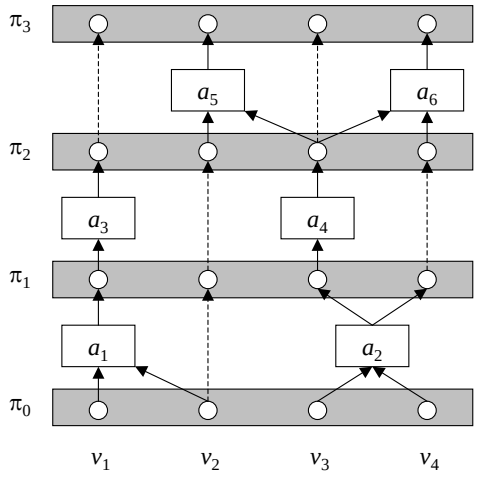

Fig. 2. Plans and action instances. Each state characterizes the values of four variables v_1, v_2, v_3 and v_4. States are changed by application of action instances

domain of an action need to be affected by the action. The actions a_5 and a_6 illustrate that concurrent actions may have overlapping domains. ■

[5] Note that since $ran_{Var}(a) \subseteq dom_{Var}(a)$, this requirement excludes overlapping ranges of concurrent actions, but domains of concurrent actions are allowed to overlap as long as the values of the variables in the overlapping domains are not affected by the actions.

3 Standard Plan Diagnosis

Let us assume, for the moment, that each action instance can be viewed as an independent component of a plan. To each action instance a a health mode $m_a \in \{nor, ab\}$ can be assigned and the result is called a *qualified* plan. In establishing which part of the plan fails, we are only interested in those actions qualifies as abnormal. Therefore, we define a qualified version P_Q of a plan $P = \langle \mathcal{A}, A, < \rangle$ as a tuple $P_Q = \langle \mathcal{A}, A, <, Q \rangle$, where $Q \subseteq A$ is the subset of instances of actions qualified as abnormal (and therefore, $A - Q$ the subset of actions qualified as normal).

Since a qualification Q corresponds to assigning the health mode ab to every action in Q and since $f_a^{ab}(d_{i_1}, d_{i_2}, \ldots, d_{i_k}) = (\perp, \perp, \ldots, \perp)$ for every action $a \in Q$ with $type(a) = \alpha$, the results of anomalously executed actions are unpredictable. Note that a "normal" plan P corresponds to the qualified plan $P_\varnothing$ and furthermore that in our context "undefined" is considered to be equivalent to "unpredictable".

3.1 Qualified Plan execution

For simplicity, when a plan P is executed, we will assume that every action takes a unit of time to execute. We are allowed to observe the execution of a plan P at discrete times $t = 0, 1, 2, \ldots, k$ where k is the depth of the plan, i.e., the longest $<$-chain of actions occurring in P. Let $depth_P(a)$ be the depth of action a in plan $P = \langle \mathcal{A}, A, < \rangle$. Here, $depth_P(a) = 0$ if $\{a' \mid a' \ll a\} = \varnothing$ and $depth_P(a) = 1 + max\{depth_P(a') \mid a' \ll a\}$, else. If the context is clear, we often will omit the subscript P. We assume that the plan starts to be executed at time $t = 0$ and that concurrency is fully exploited, i.e., if $depth_P(a) = k$, then execution of a has been completed at time $t = k + 1$. Thus, all actions a with $depth_P(a) = 0$ are completed at time $t = 1$ and every action a with $depth_P(a) = k$ will be started at time k and will be completed at time $k + 1$. Note that thanks to the above specified concurrency requirement, concurrent execution of actions having the same depth leads to a well-defined result.

Let P_t denote the set of actions a with $depth_P(a) = t$, let $P_{>t} = \bigcup_{t'>t} P_{t'}$, $P_{<t} = \bigcup_{t'<t} P_{t'}$ and $P_{[t,t']} = \bigcup_{k=t}^{t'} P_k$. Execution of P on a given initial state σ_0 will induce a sequence of states $\sigma_0, \sigma_1, \ldots, \sigma_k$, where σ_{t+1} is generated from σ_t by applying the set of actions P_t to σ_t. Instead, however, of assuming total states and total state transitions, we define the (predicted) effect of the execution of plan P on a given (partial) state π at time $t \geq 0$, denoted by (π, t).

We say that $(\pi', t + 1)$ is (directly) generated by execution of P_Q from (π, t), abbreviated by $(\pi, t) \rightarrow_{Q;P} (\pi', t + 1)$, iff the following conditions hold:

1. $\pi' \restriction ran_{Var}(a) = f_a^{nor}(\pi \restriction dom_{Var}(a))$ for each $a \in P_t - Q$ such that $dom_{Var}(a) \subseteq Var(\pi)$, that is, the consequences of all actions a enabled in π can be predicted and occur in π'.[6]
2. $Var(\pi') \cap ran_{Var}(a) = \varnothing$ for each $a \in Q \cap P_t$, since the result of executing an abnormal action cannot be predicted (even if such an action is enabled in π);

[6] An action a is enabled in a state π if $dom_{Var}(a) \subseteq Var(\pi)$.

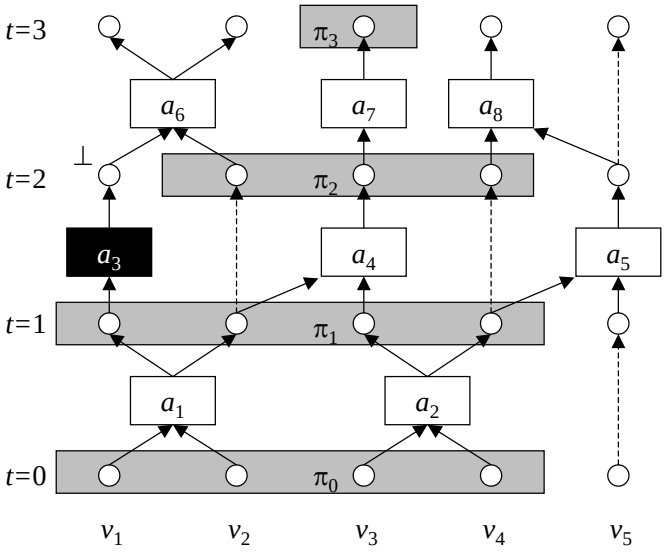

Fig. 3. Plan execution with abnormal actions

3. $Var(\pi') \cap ran_{Var}(a) = \varnothing$ for each $a \in P_t$ with $dom_{Var}(a) \not\subseteq Var(\pi)$, that is, even if an action a is enabled in (the complete state) σ_t, if a is not enabled in $\pi \sqsubseteq \sigma_t$, the result is not predictable and therefore does not occur in π', since it is not possible to predict the consequences of actions that depend on values not defined in π.
4. $\pi'(i) = \pi(i)$ for each $v_i \notin ran_{Var}(P_t)$, that is, the value of any variable not occurring in the range of an action in P_t should remain unchanged. Here, $ran_{Var}(P_t)$ is a shorthand for the union of the sets $ran_{Var}(a)$ with $a \in P_t$.

For arbitrary values of $t \leq t'$ we say that (π', t') is (directly or indirectly) generated by execution of P_Q from (π, t), denoted by $(\pi, t) \rightarrow^*_{Q;P} (\pi', t')$, iff the following conditions hold:

1. if $t = t'$ then $\pi' = \pi$;
2. if $t' = t + 1$ then $(\pi, t) \rightarrow_{Q;P} (\pi', t')$;
3. if $t' > t + 1$ then there must exists some state $(\pi'', t' - 1)$ such that $(\pi, t) \rightarrow^*_{Q;P} (\pi'', t' - 1)$ and $(\pi'', t' - 1) \rightarrow_{Q;P} (\pi', t')$.

Note that $(\pi, t) \rightarrow^*_{\varnothing;P} (\pi', t')$ denotes the normal execution of a normal plan $P_\varnothing$. Such a normal plan execution will also be denoted by $(\pi, t) \rightarrow^*_P (\pi', t')$.

Example 3. Figure 3 gives an illustration of an execution of a plan with abnormal actions. Suppose action a_3 is abnormal and generates a result that is unpredictable ($\perp$). Given the qualification $Q = \{a_3\}$ and the partially observed state π_0 at time point $t = 0$, we predict the partial states π_i as indicated in Figure 3, where $(\pi_0, t_0) \rightarrow^*_{Q;P} (\pi_i, t_i)$ for $i = 1, 2, 3$. Note that since the value of v_1 and of v_5 cannot be predicted at time $t = 2$, the result of action a_6 and of action a_8 cannot be predicted and π_3 contains only the value of v_3. ■

3.2 Diagnosis

Suppose now that we have a (partial) observation $obs(t) = (\pi, t)$ of the state of the world at time t and an observation $obs(t') = (\pi', t')$ at time $t' > t \geq 0$ during the execution of the plan P. We would like to use these observations to infer the health states of the actions occurring in P. Assuming a normal execution of P, we can (partially) predict the state of the world at a time point t' given the observation $obs(t)$: if all actions behave normally, we predict a partial state $\pi'_{\varnothing}$ at time t' such that $obs(t) \rightarrow^*_P (\pi'_{\varnothing}, t')$.

There does not need to be a strict correspondence between the variables *predicted at time* t' and the variables *observed at time* t'. That is, $Var(\pi')$ and $Var(\pi'_{\varnothing})$ need not to be identical sets. This means that to check whether the predicted state matches the observed state at time t', we have to verify whether the variables occurring in both $Var(\pi')$ and $Var(\pi'_{\varnothing})$ have identical values, that is whether $\pi'(j) = \pi'_{\varnothing}(j)$ holds for all $v_j \in Var(\pi') \cap Var(\pi'_{\varnothing})$. Therefore, these states match exactly if they are compatible i.e. $\pi' \approx \pi'_{\varnothing}$ holds.[7]

If this is not the case, the execution of some action instances must have gone wrong and we have to determine a qualification Q such that the predicted state π'_Q derived using Q is compatible with π'. Hence, we have the following straight-forward extension of the diagnosis concept in MBD to plan diagnosis (cf. [5]):

Definition 1. *Let $P = \langle \mathcal{A}, A, < \rangle$ be a plan with observations $obs(t) = (\pi, t)$ and $obs(t') = (\pi', t')$, where $t < t' \leq depth(P)$ and let $obs(t) \rightarrow^*_{Q;P} (\pi'_Q, t')$ be a derivation assuming a qualification Q.*

Then Q is said to be a plan diagnosis *of $\langle P, obs(t), obs(t') \rangle$ iff $\pi' \approx \pi'_Q$.*

Example 4. Consider again Figure 3 and suppose that we did not know that action a_3 was abnormal and that we observed $obs(0) = ((d_1, d_2, d_3, d_4), 0)$ and $obs(3) = ((d'_1, d'_3, d'_5), 3)$. Using the normal plan derivation relation starting with $obs(0)$ we will predict a state $\pi'_{\varnothing}$ at time $t = 3$ where $\pi'_{\varnothing} = (d''_1, d''_2, d''_3)$. If everything is ok, the values of the variables predicted as well as observed at time $t = 3$ should correspond, i.e. we should have $d'_j = d''_j$ for $j = 1, 3$. If, for example, only d'_1 would differ from d''_1, then we could qualify a_6 as abnormal, since then the predicted state at time $t = 3$ using $Q = \{a_6\}$ would be $\pi'_Q = (d''_3)$ and this partial state agrees with the predicted state on the value of v_3. ■

Note that for all variables in $Var(\pi') \cap Var(\pi'_Q)$, the qualification Q provides an *explanation* for the observation π' made at time point t'. Hence, for these variables the qualification provides an *abductive diagnosis* [4] for the normal observations. For all observed variables in $Var(\pi') - Var(\pi'_Q)$, no value can be predicted given the qualification Q. Hence, by declaring them to be unpredictable, possible conflicts with respect to these variables if a normal execution of all actions is assumed, are resolved. This corresponds with the idea of a *consistency-based diagnosis* [12].

The following observation shows that we might easily trivialize plan diagnoses:

Observation 1 *If $Q \subset A$ is a plan diagnosis of $\langle P, obs(t), obs(t') \rangle$, then every superset $Q' \supseteq Q$ is also a plan diagnosis and in particular A is always a plan diagnosis.*

[7] See the definition in the preliminaries.

The reason is that *(i)* $Q' \supseteq Q$ implies $\pi'_{Q'} \sqsubseteq \pi'_Q$ where $\pi'_{Q'}$ and π'_Q are the predicted states using the qualifications Q and Q', respectively and *(ii)* $\pi'_Q \approx \pi'$ and $\pi'_{Q'} \sqsubseteq \pi'_Q$, using the definition of $\approx$, immediately imply that $\pi'_{Q'} \approx \pi'$, i.e., Q' is a diagnosis as well. Since in particular $A \supseteq Q$ for every qualification Q, A is a diagnosis whenever there has been found any diagnosis.

Clearly then, the smaller a diagnosis is, the more values it will predict that are also actually observed in the resulting plan state. This, like in MBD, is a reason for us to prefer *subset-minimal* diagnoses and especially *minimum* diagnoses among the set of minimal diagnoses.

But there is a caveat: a minimum diagnosis only minimizes the number of abnormal actions to explain deviations; as important however for a diagnosis might be its *information content*, i.e. the exactness it provides in predicting the values of the variables occurring in the observed state π'. This means that besides *minimizing* the cardinality of abnormalities another criterion could be *maximizing* the exactness of the similarity by maximizing $|Var(\pi') \cap Var(\pi'_Q)|$ i.e. maximizing the number of variables having the same value in the predicted state and the observed state. Therefore, besides a minimum diagnosis we also define the notion of a *maximum informative diagnosis*:

Definition 2. *Given plan observations $\langle P, (\pi, t), (\pi', t')\rangle$, a qualification Q is said to be a* minimum plan diagnosis *if for every plan diagnosis Q' it holds that $|Q| \leq |Q'|$.*

Q is said to be a maximum informative plan-diagnosis *iff for all plan diagnoses Q^*, it holds that $|Var(\pi') \cap Var(\pi'_Q)| \geq |Var(\pi') \cap Var(\pi'_{Q^*})|$.*

Note that for every maximum informative diagnosis Q we have $Var(\pi') \cap Var(\pi'_Q) \subseteq Var(\pi') \cap Var(\pi'_\varnothing)$, where $obs(t) \rightarrow^*_{\varnothing;P} (\pi'_\varnothing, t')$ is the partial state derivation assuming a *normal plan* execution.

Also note that every maximum informative diagnosis is a minimal diagnosis. So both minimum plan diagnoses and maximum informative plan diagnoses are the result of different criteria for selecting minimal diagnoses, as the following example shows:

Example 5. To illustrate the difference between minimum plan diagnosis and maximum informative diagnosis, consider again the plan execution depicted in Figure 3. Given $obs(0)$ and $obs(3)$ and a deviation in the value of v_2 at time $t = 3$, there are three possible minimum diagnoses: $D_1 = \{a_1\}$, $D_2 = \{a_3\}$ and $D_3 = \{a_6\}$. D_2 and D_3 are also maximum-informative diagnoses. ■

4 Causes of plan-execution failures

Unlike in classical MBD, minimum diagnosis and maximum-informative diagnosis need not provide the best explanation for the differences between observed effects of a plan execution and the predicted effects. The reason is that often in a plan instances of actions do not fail independently. For example, suppose that we have a plan for carrying luggage from a depot to a number of waiting planes. Such a plan might contain several instances of a drive action pertaining to the same carrier controlled by an agent. Suppose that an instance a_i of some drive action (type) α behaves abnormally because

of malfunctioning of the carrier. Then it is reasonable to assume that other instances a_j of the same drive action that occur in the plan *after* a_i can be predicted to behave abnormally, too. Another possibility is that a number of instances of actions is related to the malfunctioning of an *agent* executing several actions in the plan. For example, in the luggage example, the carrier is controlled by a driving agent. If this agent itself is not functioning well, all driving actions as well as loading and unloading actions might be affected.

Such dependencies between action instances and between agent health states and action instances imply that sometimes qualifying an instance of an action as being abnormal implies that other instances of actions must be qualified a being abnormal, too. Minimum and information-maximum diagnosis do not take into account these dependencies between action failures. Therefore, we must take into consideration the underlying *causes* of a plan-execution failure.

4.1 Causal Rules

To be able to include a malfunctioning of an executing agent as a possible cause, we will consider a plan together with its executing agent as the system to be diagnosed. Here, an agent will be simply represented by a set H of specific health states. To identify causes of action failures, we use a set R of *causal rules* in combination with plan diagnosis. A causal rule is a rule that can appear in the following forms:

- $(\alpha_1, \alpha_2, \ldots, \alpha_k) \rightarrow \alpha_{k+1}$, where $k \geq 1$ and for $i = 1, 2 \ldots, k+1$, $\alpha_i \in \mathcal{A}$ are action types. This type of rule relates the occurrence of a set of failed actions to the occurrence of a failed action implied by them. The intuitive meaning of these rules is that if during plan execution there are, for $i = 1, \ldots, k$, action instances a_i of type α_i that have been qualified as abnormal up to time t, then it is inferred that from time $t+1$ on all instances of actions of type α_{k+1} will behave abnormally, too.
- $(h; \alpha_1, \alpha_2, \ldots, \alpha_k) \rightarrow \alpha_{k+1}$, where $k \geq 0$, $h \in H$ is a health state $(h \neq nor)$ of the plan executing agent and, for $i = 1, 2 \ldots, k+1$, $\alpha_i \in \mathcal{A}$ are action types. This type of rule relates the occurrence of an agent abnormality h and a set of action abnormalities occurring at time t to the inference of a failed action at time $t+1$. The intuitive meaning of such a rule is that if during plan execution at some time $t' \leq t+1$ the agent operates in some abnormal health states h and, for $i = 1, 2, \ldots, k$, there are action instances a_i of type α_i that have been qualified as abnormal up to time t, then it is inferred that from time $t+1$ on all instances of actions of type α_{k+1} that occur in the plan will behave abnormally, too.[8] If $k = 0$, this rule establishes a health state as a single cause for action failure.

The intuitive idea behind a causal diagnosis is to be able to explain a given plan diagnosis Q by a (usually smaller) set of qualifications (causes) Q' together with some health state h of the agent established at time t using the set of causal rules R. Using

[8] We allow abnormal health states to be detected at the same time that abnormal action consequences are generated.

such a pair consisting of a health state and a qualification should enable us to generate, using the rules in R, a set containing Q.

To define the effect of applying R to a set of (unique) instances of actions occurring in a plan, we first construct the set $inst(R)$ of instance of actions with respect to given plan $P = \langle \mathcal{A}, A, < \rangle$ as follows:

- For every rule r of the form $(\alpha_1, \alpha_2, \ldots, \alpha_k) \rightarrow \alpha_{k+1} \in R$, $inst(R)$ contains an instance $(a_{i_1}, a_{i_2}, \ldots, a_{i_k}) \rightarrow a_{i_{k+1}}$ of r whenever there exists a $t \geq 0$ such that $\{a_{i_1}, a_{i_2}, \ldots, a_{i_k}\} \subseteq P_{\leq t}$ and $a_{i_{k+1}} \in P_{>t}$.
- For every rule r of the form $(h; \alpha_1, \alpha_2, \ldots, \alpha_k) \rightarrow \alpha_{k+1} \in R$, $inst(R)$ contains the instances $(h; a_{i_1}, a_{i_2}, \ldots, a_{i_k}) \rightarrow a_{i_{k+1}}$, whenever there exists a $t \geq 0$ such that $\{a_{i_1}, a_{i_2}, \ldots, a_{i_k}\} \subseteq P_{\leq t}$ and $a_{i_{k+1}} \in P_{>t}$.

For each $r \in inst(R)$, let $ante(r)$ denote the antecedent of r and $hd(r)$ denote the head of r. Furthermore, let $Ab \subseteq \{h\}$ be a set containing an abnormal agent health state h or be equal to the empty set (signifying a normal state of the agent) and let $Q \subseteq A$ be a qualification of instances of actions. We can now define a causal consequence of a qualification Q and a health state Ab using R as follows:

Definition 3. *An instance $a \in A$ is a causal consequence of a qualification $Q \subset A$ and the health state Ab using the causal rules R if*

1. $a \in Q$ *or*
2. *there exists a rule $r \in inst(R)$ such that* (i) *for each $a_i \in ante(r)$ either a_i is a causal consequence of Q or $a_i \in Ab$, and* (ii) $a = hd(r)$.

The set of causal consequences of Q using R and Ab is denoted by $C_{R,Ab}(Q)$.

We have a simple characterization of the set of causal consequences $C_{R,Ab}(Q)$ of a qualification Q and a health state Ab using a set of causal rules R:

Observation 2 $C_{R,Ab}(Q) = Cn_A(inst(R) \cup Q \cup Ab)$.

Here, $Cn_A(X)$ restricts the set $Cn(X)$ of classical consequences of a set of propositions X to the consequences occurring in A. To avoid cumbersome notation, we will omit the subscripts R and Ab from the operator C and use $C(Q)$ to denote the set of consequences of a qualification Q using a health state Ab and a set of causal rules R. We say that a qualification Q is *closed* under the set of rules R and an agent health state Ab if $Q = C(Q)$, i.e, Q is saturated under application of the rules R.

Proposition 1. *The operator C satisfies the following properties:*

1. *(inclusion): for every $Q \subseteq A$, $Q \subseteq C(Q)$*
2. *(idempotency): for every $Q \subseteq A$, $C(Q) = C(C(Q))$*
3. *(monotony): if $Q \subseteq Q' \subseteq A$ then $C(Q) \subseteq C(Q')$*

Proof. Note that $C(Q) = Cn(inst(R) \cup Q \cup Ab) \cap A$. Hence, monotony and inclusion follow immediately as a consequence of the monotony and inclusion of Cn. Monotony and inclusion imply $C(Q) \subseteq C(C(Q))$. To prove the reverse inclusion, let $Cn^*(Q) = Cn(instr(R) \cup Q \cup Ab)$. Then by inclusion and idempotency of Cn we have $C(C(Q)) = Cn^*(C(Q)) \cap A \subseteq Cn^*(Cn^*(Q)) \cap A = Cn^*(Q) \cap A = C(Q)$. $\square$

Thanks to Proposition 1 we conclude that every qualification can be easily extended to a closed set $C(Q)$ of qualifications. Due to the presence of causal rules, we require every diagnosis Q to be closed under the application of rules, that is in the sequel we restrict diagnoses to closed sets $Q = C(Q)$.

We define a causal diagnosis as a qualification Q such that its set of consequences $C(Q)$ constitutes a diagnosis:

Definition 4. *Let $P = \langle \mathcal{A}, A, < \rangle$ be a plan, R a set of causal rules and let $obs(t)$ and $obs(t')$ be two observations with $t < t'$. Then a qualification $Q \subseteq A$ is a causal diagnosis of $(P, obs(t), obs(t'))$ if $C(Q) \cap P_{[t;t']}$ is a diagnosis of $(P, obs(t), obs(t'))$.*

Like we defined a minimum diagnosis, we now define two kinds of minimum causal diagnoses: a minimum causal *set* diagnosis and a minimum causal *effect* diagnosis:

Definition 5. *Let $P = \langle \mathcal{A}, A, < \rangle$ be a plan and $obs(t)$ and $obs(t')$ with $t < t'$ be two observations.*

1. *A* minimum causal set diagnosis *is a causal diagnosis Q such that $|Q| \leq |Q'|$ for every causal diagnosis Q' of P;*
2. *A* minimum causal effect diagnosis *is a causal diagnosis Q such that $|C(Q)| \leq |C(Q')|$ for every causal diagnosis Q'.*

Maximum informative causal set and maximum informative causal effect diagnoses are defined completely analogous to the previous definitions using standard diagnosis.

The relationships between the different diagnostic concepts we have distinguished is partially summarized in the following proposition:

Proposition 2. *Let $P = \langle \mathcal{A}, A, < \rangle$ be a plan and $obs(t)$ and $obs(t')$ with $t < t'$ be two observations.*

1. *$|Q| \leq |Q'|$ for every minimum causal set diagnosis Q and minimum closed diagnosis Q' of P;*
2. *$|Q| \leq |Q'|$ for every minimum causal effect diagnosis Q and minimum closed diagnosis Q' of P*

Proof. Both properties follow immediately from the definitions and the inclusion property of C. □

4.2 Causal diagnoses and Prediction

Except for playing a role in establishing causal *explanations* of observations, (causal) diagnoses also can play a significant role in the *prediction* of future results (states) of the plan or even the attainability of the goals of the plan. First of all, we should realize that a diagnosis can be used to enhance observed state information as follows: Suppose that Q is a causal diagnosis of a plan P based on the observations $obs(t)$ and $obs(t')$ for some $t < t'$, let $obs(t) \rightarrow^*_{C(Q);P} (\pi'_Q, t')$ and let $obs(t') = (\pi', t')$. Since $C(Q)$ is a diagnosis, π' and π'_Q are compatible states. Hence, we can combine the information contained in both partial states by merging them into a new partial state $\pi'_\sqcup = \pi'_Q \sqcup \pi'$.

This latter state can be seen as the partial state that can be obtained by direct observation at time t' as well as by making use of previous observations at time t and diagnostic information.

In the same way, we can use this information and the causal consequences $C(Q)$ to derive a prediction of the partial states derivable at a time $t'' > t'$:

Definition 6. *Let Q be a causal diagnosis of a plan P based on the observations (π, t) and (π', t') where $t < t'$. Furthermore, let $obs(t) \rightarrow^*_{C(Q);P} (\pi'_Q, t')$ and let $obs(t') = (\pi', t')$. Then, for some time $t'' > t'$, (π'', t'') is the partial state predicted using Q and the observations if $(\pi'_Q \sqcup \pi', t') \rightarrow^*_{C(Q);P} (\pi'', t'')$.*

In particular, if $t'' = depth(P)$, i.e., the plan has been executed completely, we can predict the values of some variables that will result from executing P and we can check which goals $g \in G$ will still be achieved by the execution of the plan, based on our current knowledge. That is, we can check for which goals $g \in G$ it holds that $\tau \models g$. So causal diagnosis might also help in evaluating which goals will be affected by failing actions.

4.3 Complexity issues

It is well-known that the diagnosis problem is computationally intractable. The decision forms of both consistency-based and abductive based diagnosis are NP-hard ([2]). It is easy to see that standard plan diagnosis has the same order of complexity. Concerning (minimal) causal diagnoses, we can show that they are not more complex than establishing plan diagnoses if the latter problem is NP-hard. The reason is that in every case the verification of Q' being a causal diagnosis is as difficult as verifying a plan diagnosis under the assumption that the set $inst_P(R)$ is polynomially bounded in the size $||P||$ of the plan P.[9] Also note that subset minimality (under a set of rules $inst(R)$ of a set of causes can be checked in polynomial time.

5 Related research

In this section we briefly discuss some other approaches to plan diagnosis. Like we use MBD as a starting point to plan diagnosis, Birnbaum et al. [1] apply MBD to *planning agents* relating health states of agents to *outcomes* of their planning activities, but not taking into account faults that can be attributed to actions occurring in a plan as a separate source of errors. However, instead of focusing upon the relationship between agent properties and outcomes of plan executions, we take a more detailed approach, distinguishing two separate sources of errors (actions and properties of the executing agents) and focusing upon the detection of anomalies during the plan execution. This enables us to predict the outcomes of a plan on beforehand instead of using them only as observations.

[9] The reason is that computing consequences of Horn-theories can be achieved in a time linear in the size of $inst_P(R)$.

de Jonge et al. [6] propose another approach that directly applies model-based diagnosis to plan execution. Their paper focuses on agents each having an individual plan, and where conflicts between these plans may arise (e.g. if they require the same resource). Diagnosis is applied to determine those factors that are accountable for *future* conflicts. The authors, however, do not take into account dependencies between health modes of actions and do not consider agents that collaborate to execute a common plan.

Kalech and Kaminka [9, 10] apply *social diagnosis* in order to find the cause of an anomalous plan execution. They consider hierarchical plans consisting of so-called *behaviors*. Such plans do not prescribe a (partial) execution order on a set of actions. Instead, based on its observations and beliefs, each agent chooses the appropriate behavior to be executed. Each behavior in turn may consist of primitive actions to be executed, or of a set of other behaviors to choose from. Social diagnosis then addresses the issue of determining what went wrong in the joint execution of such a plan by identifying the disagreeing agents and the causes for their selection of incompatible behaviors (e.g., belief disagreement, communication errors). This approach might complement our approach when conflicts not only arise as the consequence of faulty actions, but also as the consequence of different selections of sub-plans in a joint plan.

Lesser et al. [3, 8] also apply diagnosis to (multi-agent) plans. Their research concentrates on the use of a *causal model* that can help an agent to refine its initial diagnosis of a failing *component* (called a *task*) of a plan. As a consequence of using such a causal model, the agent would be able to generate a new, situation-specific plan that is better suited to pursue its goal. While their approach in its ultimate intentions (establishing anomalies in order to find a suitable plan repair) comes close to our approach, their approach to diagnosis concentrates on specifying the exact causes of the failing of one single *component* (task) of a plan. Diagnosis is based on observations of a component without taking into account the consequences of failures of such a component w.r.t. the remaining plan. In our approach, instead, we are interested in applying MBD-inspired methods to *detect* plan failures. Such failures are based on observations during plan execution and may concern individual components of the plan, but also agent properties. Furthermore, we do not only concentrate on failing components themselves, but also on the consequences of these failures for the future execution of plan elements.

6 Conclusion

We have adapted model-based agent diagnosis to the diagnosis of plans and we have pointed out some differences with the classical approaches to diagnosis. We distinguished two types of diagnosis: minimum plan diagnosis and maximum informative diagnosis to identify (*i*) minimum sets of anomalously executed actions and (*ii*) maximum informative (w.r.t. to predicting the observations) sets of anomalously executed actions. Assuming that a plan is carried out by a single agent, anomalously executed action can be correlated if the anomaly is caused by some malfunctions in the agent. Therefore, (*iii*) causal diagnoses have been introduced and we have extended the diagnostic theory enabling the prediction of future failure of actions.

Current work can be extended in several ways. We mention two possible extensions: First of all, we could improve the diagnostic model of the executing agent. The

causal diagnoses are based on the assumption that the agent enters an abnormal state at some time point and stays in that state until the agent is repaired. In our future work we wish to extend the model such that the agent might evolve through several abnormal states. The resulting model will be related diagnosis in Discrete Event Systems [7, 11]. Moreover, we intend to investigate plan repair in the context of the agent's current (abnormal) state. Secondly, we would like to extend the diagnostic model with sequential observations and iterative diagnoses. Here, we would like to consider the possibilities of diagnosing a plan if more than two subsequent observations are made, the best way to detect errors in such cases and the construction of enhanced prediction methods.

Acknowledgements This research is supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Dutch Ministry of Economic Affairs.

References

1. L. Birnbaum, G. Collins, M. Freed, and B. Krulwich. Model-based diagnosis of planning failures. In *AAAI 90*, pages 318–323, 1990.
2. T. Bylander, D. Allemang, M. C. Tanner, and J. R. Josephson. The computational complexity of abduction. *Artif. Intell.*, 49(1-3):25–60, 1991.
3. N. Carver and V.R. Lesser. Domain monotonicity and the performance of local solutions strategies for cdps-based distributed sensor interpretation and distributed diagnosis. *Autonomous Agents and Multi-Agent Systems*, 6(1):35–76, 2003.
4. L. Console and P. Torasso. Hypothetical reasoning in causal models. *International Journal of Intelligence Systems*, 5:83–124, 1990.
5. L. Console and P. Torasso. A spectrum of logical definitions of model-based diagnosis. *Computational Intelligence*, 7:133–141, 1991.
6. F. de Jonge and N. Roos. Plan-execution health repair in a multi-agent system. In *PlanSIG 2004*, 2004.
7. R. Debouk, S. Lafortune, and D. Teneketzis. Coordinated decentralized protocols for failure diagnosis of discrete-event systems. *Journal of Discrete Event Dynamical Systems: Theory and Application*, 10:33–86, 2000.
8. B. Horling, B. Benyo, and V. Lesser. Using Self-Diagnosis to Adapt Organizational Structures. In *Proceedings of the 5th International Conference on Autonomous Agents*, pages 529–536. ACM Press, 2001.
9. M. Kalech and G. A. Kaminka. On the design ov social diagnosis algorithms for multi-agent teams. In *IJCAI-03*, pages 370–375, 2003.
10. M. Kalech and G. A. Kaminka. Diagnosing a team of agents: Scaling-up. In *AAMAS 2004*, 2004.
11. Y. Pencolé and M. Cordier. A formal framework for the decentralised diagnosis of large scale discrete event systems and its application to telecommunication networks. *Artif. Intell.*, 164(1-2):121–170, 2005.
12. R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.