Diagnosis of Plans and Agents

Nico Roos[1] and Cees Witteveen[2]

[1] Department 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

Abstract. We discuss the application of Model-Based Diagnosis in (agent-based) planning. Here, a plan together with its executing agent is considered as a system to be diagnosed. It is assumed that the execution of a plan can be monitored by making partial observations of the results of actions. These observations are used to explain the observed deviations from the plan by 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: such failings may occur as the result of malfunctioning of the executing agent or may be caused by 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 determine the underlying causes of plan failing and to predict future anomalies in the execution of actions.

1 Introduction

The well-known quote: *"No plan survives its first contact with the enemy"* should remind us that *diagnosis* constitutes an unavoidable part of the plan execution process.[1] Here, plan diagnosis might refer to quite different aspects of a failing plan in execution. Since there is a huge number of potential factors that might influence, or even prevent, correct plan execution, it is not surprising that current approaches to plan diagnosis are rather diverse.

The aim of this paper is to adapt and extend a classical Model-Based Diagnosis (MBD) approach to the diagnosis of plans. First, we will first show how a plan consisting of a partially ordered set of actions can be viewed as a system to be diagnosed by proposing an object oriented description of an action's behavior. Given this view, a diagnosis can be established using *partial observations* of a plan in progress.

Second, we introduce the concept of a *causal diagnosis*. Traditional MBD focuses on minimal diagnosis based on the intuitively acceptable assumption that components qualified as abnormal are failing *independently* from each other. However, as soon as *dependencies* exist between such components, the choice for minimal diagnoses cannot

[1] The quote is attributed to the Prussian Field Marshall Von Moltke.

M. Pěchouček, P. Petta, and L.Z. Varga (Eds.): CEEMAS 2005, LNAI 3690, pp. 357–366, 2005.

be justified. As we will argue, the existence of dependencies between failing actions in a plan is often the rule instead of an exception.

Finally, we will introduce causal rules and causal diagnoses to predict future failings of actions.

Related Work. We briefly discuss some other approaches to plan diagnosis. Similar to our use of 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 they do not take into account faults that can be attributed to actions occurring in a plan as a separate source of errors.

de Jonge et al. [5] 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).

Lesser et al. [2,8] 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. While their approach in its ultimate intentions comes close to our approach, their approach to diagnosis concentrates on specifying the exact causes of the failing of one single *component* (tasks) of a plan. Diagnosis is based on observations of a single component without taking into account the consequences of failures of such a component w.r.t. the remaining plan.

Paper Outline. 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. Section 5 concludes the paper.

2 Preliminaries

Model Based Diagnosis. Classical Model-Based Diagnosis (MBD) [3,4,12] uses a model of a system to identify causes of discrepancies between the observed behavior of the system and the behavior predicted by the model. The model that is applied consists of a set $Comp$ of components, a set M_c of health modes for each component $c \in Comp$, and a specification of a component's behavior given a health mode. 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 abnormal to be minimized.

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 discuss our *object-* or *resource-based* view on the world, assuming that for the planning problem at hand, the world can be simply described by a set $Obj = \{o_1, o_2, \ldots, o_n\}$ of objects, their respective *value domains* S_i and and their (current) values $s_i \in S_i$.[2] A *state of the world* σ then is an element of $S_1 \times S_2 \times \ldots \times S_n$. It will not always be possible to give a complete state description. Therefore, we introduce a *partial state* as an element $\pi \in S_{i_1} \times S_{i_2} \times \ldots \times S_{i_k}$, where $1 \leq k \leq n$ and $1 \leq i_1 < \ldots < i_k \leq n$. We use $O(\pi)$ to denote the set of objects $\{o_{i_1}, o_{i_2}, \ldots, o_{i_k}\} \subseteq Obj$ specified in such a state π. The value s_j of object $o_j \in O(\pi)$ in π will be denoted by $\pi(j)$. The value of an object $o_j \in Obj$ not occurring in a partial state π is said to be unknown (or unpredictable) in π, denoted by $\perp$. Partial states can be ordered with respect to their information content: π is said to be contained in π', denoted by $\pi \sqsubseteq \pi'$, iff $O(\pi) \subseteq O(\pi')$ and $\pi'(j) = \pi(j)$ for every $o_j \in O(\pi)$. We say that two partial states π, π' are *equivalent* modulo a set of objects O, denoted by $\pi =_O \pi'$, if for every $o_j \in O$, $\pi(j) = \pi'(j)$. Finally, we define the partial state π restricted to a given set O, denoted by $\pi \restriction O$, as the state $\pi' \sqsubseteq \pi$ such that $O(\pi') = O \cap O(\pi)$.

Goals. An (elementary) goal g of an agent specifies a set of 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 $S_{i_1} \times \ldots \times S_{i_k}$ 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 $(v_{i_1}, v_{i_2}, \ldots, v_{i_k})$ such that $(v_{i_1}, v_{i_2}, \ldots v_{i_k}) \sqsubseteq \pi$. 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 $O_\alpha \subseteq Obj$ by other values, dependent upon the values of another set $O'_\alpha \supseteq O_\alpha$ of objects. Hence, every action scheme α can be modeled as a (partial) function $f_\alpha : S_{i_1} \times \ldots \times S_{i_k} \rightarrow S_{j_1} \times \ldots \times S_{j_l}$, where $1 \leq i_1 < \ldots < i_k \leq n$ and $\{j_1, \ldots, j_l\} \subseteq \{i_1, \ldots, i_k\}$. The objects whose value domains occur in $dom(f_\alpha)$, the *input resources* of α, will be denoted by $dom_O(\alpha) = \{o_{i_1}, \ldots, o_{i_k}\}$ and, likewise $ran_O(\alpha) = \{o_{j_1}, \ldots, o_{j_l}\}$ denotes the *output resources* of α. Note that $ran_O(\alpha) \subseteq dom_O(\alpha)$. This functional specification f_α constitutes the *normal* behavior of the action scheme, denoted by f_α^{nor}.

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. To keep the discussion simple, in the sequel we only consider two health modes, the normal behavior mode: nor, and the

[2] In contrast to the conventional approach to state-based planning, cf. [7].

most general abnormal behavior mode: ab. The most general abnormal behavior of action α is specified by the function f_α^{ab}, where $f_\alpha^{ab}(s_{i_1}, s_{i_2}, \ldots, s_{i_k}) = (\bot, \bot, \ldots, \bot)$.[3]

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 $type(a_i) = \alpha \in \mathcal{A}$, a_i is said to be of type α. If the context permits we will use "actions" and "instances of actions" interchangeably.

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 precedence 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 $ran_O(a) \cap dom_O(a') \neq \varnothing$ then $a < a'$ or $a' < a$.[4]

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

Example 1. Figure 1 gives an illustration of a plan. Arrows relate the objects an action uses as inputs and produces as its outputs to the action itself. In this plan, the dependency relation is specified as $a_1 \ll a_3$, $a_1 \ll a_4$, $a_2 \ll a_4$, $a_2 \ll a_5$, $a_4 \ll a_7$, $a_5 \ll a_8$ and $a_4 \ll a_6$. Note that the last dependency has to be included because a_6 changes the value of o_2 needed by a_4. The actions a_4, a_5 and a_6 show that not every object occurring in the domain of an action needs to be affected by the action. ■

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 qualified 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 $f_a^{ab}(s_{i_1}, s_{i_2}, \ldots, s_{i_k}) = (\bot, \bot, \ldots, \bot)$ for every action $a \in Q$ with $type(a) = \alpha$, the results of anomalously executed actions are unpredictable.

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

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

[4] Note that since $ran_O(a) \subseteq dom_O(a)$, this requirement excludes overlapping ranges of concurrent actions, but domains of concurrent actions are allowed to overlap as long as the values of the object in the overlapping domains are not affected by the actions.

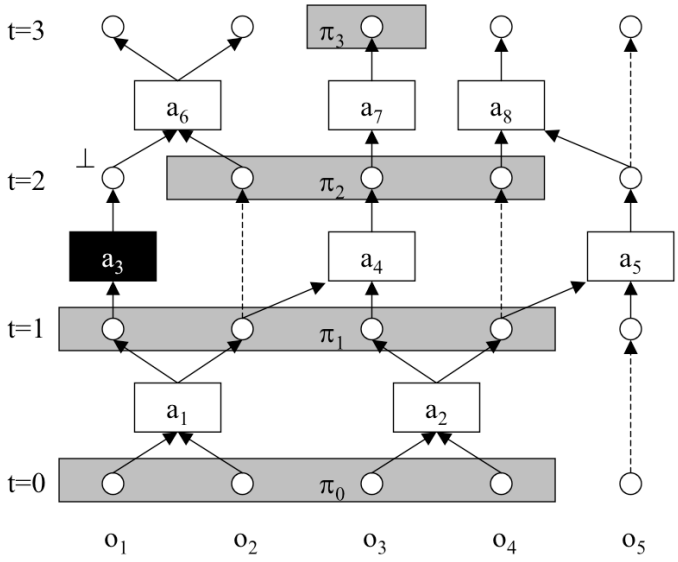

Fig. 1. Plan execution with abnormal actions

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 $P = \langle \mathcal{A}, A, < \rangle$.[5] 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. Generalizing to partial states and transitions from partial states, 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. $\pi' \restriction ran_O(a) = f_a^{nor}(\pi \restriction dom_O(a))$ for each $a \in P_t - Q$ such that $dom_O(a) \subseteq O(\pi)$, that is, the consequences of all actions a enabled in π can be predicted and occur in π'.[6]

[5] 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.

[6] An action a is enabled in a state π if $dom_O(a) \subseteq O(\pi)$.

2. $O(\pi') \cap ran_O(a) = \varnothing$ 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);
3. $O(\pi') \cap ran_O(a) = \varnothing$ for each $a \in P_t$ with $dom_O(a) \not\subseteq O(\pi)$, 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 $o_i \notin ran_O(P_t)$, that is, the value of any object not occurring in the range of an action in P_t should remain unchanged. Here, $ran_O(P_t)$ is a shorthand for the union of the sets $ran_O(a)$ 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 $t' \geq t + 1$ then $(\pi, t) \rightarrow_{Q;P} (\pi'', t + 1)$ 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 $P_\varnothing$.

Example 2. Figure 1 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 1, where $(\pi_0, t_0) \rightarrow^*_{Q;P} (\pi_i, t_i)$ for $i = 1, 2, 3$. Note that since the value of o_1 and of o_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 o_3. ■

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 $obs(t) \rightarrow^*_P (\pi'_\varnothing, t')$. Since we do not require observations to be made systematically, $O(\pi')$ and $O(\pi'_\varnothing)$ might only partially overlap. Therefore, if all actions are executed normally, the values of the objects that occur in both the predicted state and the observed state at time t' should match, i.e, we should have

$$\pi' =_{O(\pi') \cap O(\pi'_\varnothing)} \pi'_\varnothing.$$

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 derived using Q agrees with π'. This is nothing else then a straight-forward extension of the diagnosis concept in MBD to plan diagnosis (cf. [4]):

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 $\pi' =_{O(\pi') \cap O(\pi'_Q)} \pi'_Q$.*

So in a plan diagnosis Q the observed partial state (π') at time t' and the predicted state (π'_Q) assuming the qualification Q at time t' agree upon the values of all objects occurring in both states.

Example 3. Consider again Figure 1 and suppose that we did not know that action a_3 was abnormal and that we observed $obs(0) = ((s_1, s_2, s_3, s_4), 0)$ and $obs(3) = (s'_1, s'_3, s'_5), 3)$. Using the normal plan derivation relation starting with $obs(0)$ we will predict a state $\pi'_\varnothing$ at time $t = 3$ where $\pi'_\varnothing = (s''_1, s''_2, s''_3)$. If everything is ok, the values of the objects predicted as well as observed at time $t = 3$ should correspond, i.e. we should have $s'_j = s''_j$ for $j = 1, 3$. If, for example, only s'_1 would differ from s''_1, then we could qualify a_6 as abnormal, since then the predicted state at time $t = 3$ using $Q = \{a_6\}$ would be $\pi'_Q = (s''_3)$ and this partial state agrees with the predicted state on the value of o_3. ■

Note that for all objects in $O(\pi') \cap O(\pi'_Q)$, the qualification Q provides an *explanation* for the observation π' made at time point t'. Hence, for these objects the qualification provides an *abductive diagnosis* [3] for the normal observations. For all observed objects in $O(\pi') - O(\pi'_Q)$, no value can be predicted given the qualification Q. Hence, by declaring them to be unpredictable, possible conflicts with respect to these objects if a normal execution of all actions is assumed, are resolved. This corresponds with the idea of a *consistency-based diagnosis* [12].

If Q is a plan diagnosis of $\langle P, obs(t), obs(t')\rangle$, then every superset $Q' \supseteq Q$ is also a plan diagnosis, since in that case we have $\pi'_{Q'} \sqsubseteq \pi'_Q$ and therefore $\pi' =_{O(\pi') \cap O(\pi'_Q)} \pi'_Q$ implies $\pi' =_{O(\pi') \cap O(\pi'_{Q'})} \pi'_{Q'}$. 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 *minimum* diagnoses among the set of minimal diagnoses.

But there is a caveat: a minimum diagnosis only minimizes abnormalities 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* $|O(\pi') \cap O(\pi'_Q)|$.

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 $|O(\pi') \cap O(\pi'_Q)| \geq |O(\pi') \cap O(\pi'_{Q^*})|$.*

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 4. To illustrate the difference between minimum plan diagnosis and maximum informative diagnosis, consider again the plan execution depicted in Figure 1. Given $obs(0)$ and $obs(3)$ and a deviation in the value of o_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 these dependencies between action failures into account. Therefore, we must take into consideration the underlying *causes* of a plan-execution failure.

Causal Rules. We consider a plan together with its executing agent as the system to be diagnosed. An agent will be 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. The intuitive idea behind causal rules is that the rules enables us to predict failures of future actions given the agent's health state and a set of failed actions. A causal rule is a rule that can appear in the following forms:

> $(h; \alpha_1, \alpha_2, \ldots, \alpha_k) \rightarrow \alpha_{k+1}$, where $k \geq 0$, $h \in H$ is a health state 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 health state h and a set of action abnormalities occurring at time t to the inference of a failed action at time $t+1$. If $k = 0$ and $h \neq nor$, this rule establishes a health state as a single cause for action failure.

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 $(h; \alpha_1, \alpha_2, \ldots, \alpha_k) \rightarrow \alpha_{k+1} \in R$, $inst(R)$ 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 $a_{i_{k+1}} \in P_{>t}$.

Note that the failure of an action $a_{i_{k+1}}$ only depends on $a_{i_1}, a_{i_2}, \ldots, a_{i_k}$ if the agent is healthy: $h = nor$.

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 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.

Definition 3. *The set of a causal consequence $C_{R,h}(Q)$ of a qualification $Q \subset A$ given the health state $h \in H$ and the causal rules R is defined as:*

$$C_{R,h}(Q) = Cn_A(inst(R) \cup Q \cup \{h\}).$$

Here, the instances of causal rules are interpreted as Horn clauses, Q and $\{h\}$ as sets of atoms, and Cn denotes the logical consequence operator.

To simplify the notation, we will omit the subscripts R and h from the operator C.

Now 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 $(P, obs(t), obs(t'))$.*

Among the causal diagnoses, we distinguish *minimum* and *maximum informative* causal diagnoses. Moreover, we distinguish *closed set* causal diagnoses; i.e. causal diagnoses Q such that $C(Q) = Q$.

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 agree upon the values of all objects occurring in both states. Therefore we can combine the information contained in both partial states by merging them into a new partial state $\pi'_\sqcup = \pi'_Q \sqcup \pi'$. Here, the merge $\pi^1 \sqcup \pi^2$ of two partial states π^1 and π^2 is simply defined as the partial state π where $\pi(j) = \pi^i(j)$ iff $\pi^i(j)$ is defined for $i = 1, 2$ and undefined else. $\pi'_\sqcup$ can be seen as the partial state that can be obtained by direct observation at time t and indirectly by making use of previous observations and plan 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 times $t'' > t'$:

Definition 5. *Let Q is 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 objects 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.

5 Conclusion

We have presented a new object-oriented model to specify plans and to apply techniques developed for model-based agent 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. We intend to extend our model along three lines. First, we wish to extend the model such that the agent might evolve through several abnormal states. The resulting model will be related to diagnosis in Discrete Event Systems [6,11]. Second, we intend to investigate plan repair in the context of the agent's current (abnormal) state. Third, we would like to extend the diagnostic model with sequential observations and iterative diagnoses.

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. 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.
3. L. Console and P. Torasso. Hypothetical reasoning in causal models. *International Journal of Intelligence Systems*, 5:83–124, 1990.
4. L. Console and P. Torasso. A spectrum of logical definitions of model-based diagnosis. *Computational Intelligence*, 7:133–141, 1991.
5. F. de Jonge and N. Roos. Plan-execution health repair in a multi-agent system. In *PlanSIG 2004*, 2004.
6. 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.
7. R. E. Fikes and N. Nilsson. Strips: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 5:189–208, 1971.
8. Bryan Horling, Brett Benyo, and Victor 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 of 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, M. Cordier, and L. Rozé. Incremental decentralized diagnosis approach for the supervision of a telecommunication network. In *DX01*, 2001.
12. R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.