Belief Management using the Action History and Consistency-Based-Diagnosis

Clemens Mühlbacher and Gerald Steinbauer
* Institute for Software Technology
Graz University of Technology
Graz, Austria

Abstract

The belief of an agent describes how the agents think the world looks like. According to this belief the agent performs action towards its goals. Only if the belief is consistent with the real world the agent will achieve the goal. Otherwise the agent may perform useless or even dangerous actions. The belief of the agent is drawn from the effect of performed actions and the result of performed sensing actions. Due to unreliable action execution and faulty sensor readings the belief cannot always be kept consistent. To deal with this inconsistency we propose a diagnosis approach which uses the history of performed actions and the so called action influence. Within this paper we show how the diagnosis problem is formalized, how a conflict driven search can be performed and how the agent can draw its belief from the set of diagnosis.

1 Introduction

It is very common for agents to perform their actions according to the belief the agent has about its environment. In order to generate such a belief the agent uses the knowledge about the initial situation, the effects of performed actions and sensing results. As actions not always settles their effects as they should and sensing does not always yield the correct result the belief of the agent may not reflect the state of the environment correctly. The resulting inconsistency in the belief of the agent may hinder the agent to fulfill its task successfully.

In order to deal with an inconsistent belief different methods have been developed over the years. Many of these approaches consider a model of the possible faults of an action. Thus the agent must have knowledge about how an action can fail. Furthermore the action is only allowed to fail in that way. If an action fails in an unforeseen way the approaches can no longer come up with reasonable explanations. This is a big drawback as it is often not feasible to consider all possibilities an action can fail.

Let's consider a simple example to point out the problem with known faults. An agent can moves boxes which can contain objects. The movement of an object can fail. If the agent now moves a box it may also drop the box. But what we need to consider if the agent drops a box. (1) Only the box is dropped but the objects are still within the box. (2) All object fall out of the box, (3) .... But besides dropping the complete box the agent can just move the box in a way that some of the objects are lost in the box. To consider all these possibilities we would be forced to come up with a complex description which faults are possible.

In contrast one can simple state that the agent can move the box and can fail to achieve the effect of the movement in some way. The only things which are influence by this fault are the box and the objects within the box. Thus one could simple cover the case that after failing to move the box it could be the case that one object is on the floor but another object and the box are successful delivered.

Within this paper we show how an agent can follow this idea to draw its belief out of executed actions. To do so we will show how to formalize the reasoning with uncertain action outcomes. Furthermore we formalize a diagnosis problem to identify the failing actions. Additionally we show how a conflict driven search can be performed in order to calculate the diagnosis. With the help of this diagnosis calculation the agent can draw its belief. As the formalization poses a complex non monotonic reasoning task we show how this task can be formalized with default logic as well.

In the next section we will discuss the preliminaries how an agent can perform a reasoning with the help of an action history. The proceeding section will show how the diagnosis problem is formalized and how the agent draws its belief from the set of diagnosis. The following section discusses how one can perform a conflict driven search for the diagnosis. Afterwards we show how one can use default logic to specify the diagnosis and the reasoning about the belief within one logical framework. Followed by a discussion how this approach relates to an approach using fault modes. In Section 7 we brief discuss the relation to other belief management approaches. Finally we will conclude the paper in point out future work.

2 Preliminaries

In order to reason with a belief an agent needs to reason about performed action. One possibility is to use the situation calculus [McCarthy, 1963] to perform this reasoning. The situation calculus is a second order logical language with equality. A situation represents a sequence of performed actions. The constant S_0 is used to donate the initial situation. Additionally the function $s' = do(\alpha, s)$ is

*This work is supported by the Austrian Research Promotion Agency (FFG) under grant 843468 (Guaranteeing Service Robot Dependability During the Entire Life Cycle (GUARD)).

used to specify that action α was performed in situation s and situation s' is the result of this action.

To specify the state of the world which is changed predicate which change over time are used. These predicates are called fluents. As actions influence the world the situation calculus use successor state axioms [Reiter, 2001] to change the value of a fluent after an action was performed. The successor state axiom is defined as follows: $F(\vec{x}, do(\alpha, s)) \equiv \varphi^+(\alpha, \vec{x}, s) \vee F(\vec{x}, s) \wedge \neg\varphi^-(\alpha, \vec{x}, s)$. $\varphi^+(\alpha, \vec{x}, s)$ respectively $\varphi^-(\alpha, \vec{x}, s)$ is used to state under which condition action α sets the value of the fluent to true respectively false.

As an action α can only be performed under certain conditions the action precondition Π_α is used to express these restrictions. Finally unique name assumptions for actions $\mathcal{D}_{una}$, axioms describing the initial situation $\mathcal{D}_{S_0}$ and some foundational axioms Σ are combined with the successor state axioms $\mathcal{D}_{ssa}$ and the precondition axioms $\mathcal{D}_{ap}$ to form the so called basic action theory: $\mathcal{D} = \mathcal{D}_{ap} \cup \mathcal{D}_{ssa} \cup \mathcal{D}_{una} \cup \mathcal{D}_{S_0} \cup \Sigma$.

In order to define the impact of a sensing action within the situation calculus the predicate $SF(\alpha) \equiv \phi(s)$ is used [Scherl and Levesque, 1993]. The formula is used to constraint the situation after the action α has been performed. If the action α is performed and it yields the sensing result $\top$ $SF(\alpha)$ is asserted. If the result of the action is $\bot$ $\neg SF(\alpha)$ is asserted. These assertions of the performed actions are added to the basic action theory $\mathcal{D}$ for the reasoning process. Thus if one wants to know if a formula ψ holds in situation s one performs the following logical entailment $\mathcal{D} \cup \{SF(\alpha)\} \models \psi(s)$. Thus it follows that after performing a sensing action with the result $\top$ $SF(\alpha)$ can be concluded. Furthermore if the situation s entails $SF(\alpha)$ and the action α is executed in this situation the sensing result $\bot$ yields a contradiction. Such a contradiction is caused due to a fault in the sensing or in the execution of the action. In the next section we will discuss how to deal with such faults.

3 Consistency History Based Belief Management

To deal with faults during the action execution we will follow the idea of history based diagnosis [Gspandl *et al.*, 2011]. The idea of history based diagnosis is to alter the history of performed actions until the history is consistent again. In order to do so fault modes for each action are used. Furthermore the usage of a fault mode implies additional costs on the history. These costs represent the likelihood of the history. Histories with low costs are more likely. Thus basically a diagnosis with fault modes is performed on the history of performed actions.

As we argued above fault modes are often not feasible to derive. As such we will perform a consistency based diagnosis on the history of performed actions. For each action in the history we determine if it is faulty or not. If blamed to be faulty the action outcome is unknown and can be arbitrary. Thus this idea follows closely the idea presented in [Witteveen *et al.*, 2005].

In order to do so we define an action as failing through the predicate $fail(\alpha, s)$. The predicate $fail(\alpha, s)$ stating that action α has failed in situation s. Thus action α which was performed to produce $s' = do(\alpha, s)$ is blamed to be faulty and has basically not settled its expected effects.

As actions which are fault may cause an arbitrary effect we need to define the influence an action can have. Let's consider the example from the introduction. Moving the box influence the robot position, the box position and the position of all objects with the box. But it does not influence the color of the robot. To define this influence we define for each fluent F the predicate $\psi_\alpha^F(\vec{x}, s)$ stating if the action α influence the fluent $F(\vec{x}, s)$.

With the help of this influence we can specify the effect a faulty action has. To specify the effect a faulty action has, we need to modify the successor state axioms. In order to do so we will follow the idea of the so called generalized successor state axioms [De Giacomo *et al.*, 2001]. The successor state axiom for fluent $F(\vec{x}, do(\alpha, s))$ is now defined as follows:

$$\begin{aligned}&[\psi_\alpha^F(\vec{x}, s) \rightarrow (\neg fail(\alpha, s) \wedge \Pi_\alpha(s))] \\ \rightarrow&[F(\vec{x}, do(\alpha, s)) \leftrightarrow \varphi^+(\alpha, \vec{x}, s) \\ &\vee F(\vec{x}, s) \wedge \neg\varphi^-(\alpha, \vec{x}, s)]\end{aligned} \tag{1}$$

Thus the successor state axiom is only applicable if the action α within the history influences the fluent, $\psi_\alpha^F(\vec{x}, s)$ is true, the action is not allowed to be faulty and the precondition of the action needs to hold. Thus if the action influence the fluent and either it is fault or the precondition does not hold the fluent value becomes unknown.

To specify the influence an action can have one can follow three simple strategies:

1. The maximal effect a fault can have can be estimate. Thus $\psi_\alpha^F(\vec{x}, s)$ is a precise formula. This was the case of our running example.
2. A fault of an action influence only those fluents the action would influence in case of a successful execution. $\psi_\alpha^F(\vec{x}, s) \equiv \varphi^+(\alpha, \vec{x}, s) \vee \varphi^-(\alpha, \vec{x}, s)$. This could be the case for a goto action of a robot. This only influences the robot pose and the related fluents.
3. A fault influence the same fluents as if the action would successfully been executed, but without restriction to the parameters of the fluent. $\psi_\alpha^F(\vec{x}, s) \equiv \exists\vec{x}'.(\varphi^+(\alpha, \vec{x}', s) \vee \varphi^-(\alpha, \vec{x}', s))$. This could be the case for a stack operation the agent performs. If the action fails the positions on the stack of all objects may get influenced.

Beside the effect an action have we also needed to take the result of the sensing action into account in order to create a consistent history. In order to deal correctly with the sensing result we use the predicate $AO(s)$ to indicate if the sensing action outcome was either $\top$ or $\bot$. Thus if the action α leading to the situation $s = do(\alpha, s')$ yields $\top$ as sensing result $AO(s) \equiv \top$. All the predicates with their truth value are gathered in the set $\mathcal{D}_{AO}$.

As the agent acts in an environment with certain constraints we can place invariants which need to hold in every situation. These invariants can help to detect inconsistency but also allow one to restrict uncertainty. This is of special interest as due to fault of an action the agent may not be able to determine the truth value of certain fluents any more. Thus with the help of these invariants this uncertainty can be reduced. To define this invariant we use the predicate $BK(s) \doteq \bigwedge_{\iota_i(s) \in Invariatns} \iota_i(s)$ which states that the invariant ι_i hold for the given situation. In the remaining of the paper we assume that the initial situation is consistent with the background knowledge thus $\mathcal{D}_{S_0} \cup BK(S_0) \not\models \bot$.

With the above predicates we are able to define under which circumstances a situation s is a consistent history.

Definition 1. *A history s is a consistent history iff:*

1. $Cons(S_0) \doteq \top$
2. $$Cons(\bar{s}) \doteq \exists s.Cons(s) \wedge BK(\bar{s}) \wedge \bar{s} = do(\alpha, s) \wedge ((\neg fail(\alpha, s) \wedge \Pi_\alpha(s)) \rightarrow (AO(\bar{s}) \leftrightarrow SF(\alpha, \bar{s})))$$

The first part of the definition just state that the initial situation needs to be consistent. The second part specifies the consistency in a recursive manner. Additionally the background knowledge needs to hold in the situation. Finally the sensing effect needs to be settled if the action as not failed and the precondition of the action hold.

With the above definition of a consistency we can now define the diagnosis as follows:

Definition 2. *A diagnosis δ for a history s is set of faulty actions: $\delta = \{\langle \alpha, s\rangle | fail(\alpha, s)\}$. Furthermore for a diagnosis δ it need to hold: $\mathcal{D} \cup \mathcal{D}_{AO} \cup \{fail(\alpha, s) | \langle \alpha, s\rangle \in \delta\} \cup \{\neg fail(\alpha, s) | \langle \alpha, s\rangle \notin \delta\} \models Cons(s)$*

Following Occam's razor we will use the set Δ comprising the diagnosis with the minimal cardinality. Thus the diagnosis with the minimum number of faulty actions forms the most plausible histories for the agent.

In order to use the most plausible histories for the agent we define the consistency basic action theory (CBAT) to reason about a history with a given diagnosis: $\mathcal{D}^{cons}(\delta) = \mathcal{D}_{ap} \cup \mathcal{D}_{ssa'} \cup \mathcal{D}_{una} \cup \mathcal{D}_{S_0} \cup \Sigma \cup \mathcal{D}_{AO} \cup \{fail(\alpha, s) | \langle \alpha, s\rangle \in \delta\} \cup \{\neg fail(\alpha, s) | \langle \alpha, s\rangle \notin \delta\}$ where $\mathcal{D}_{ssa'}$ are the successor state axioms as defined above.

Finally we can define the belief of an agent as follows:

Definition 3. *An agent's belief a formula ϕ is defined as: $Belief(\phi, s) \doteq \forall \delta \in \Delta. \mathcal{D}^{cons}(\delta) \models \phi(s)$*

In the remaining of the section we will discuss some theoretical results which hold through the definition of the diagnosis and the belief.

In order to state some theoretical results we place the following assumptions.

Assumption 1. *Actions are only executable if there precondition is believed to hold. $Belief(\Pi_\alpha, s)$*

Assumption 2. *We assume that no action has a contradicting effect. $\mathcal{D}^{cons}(\{\}) \models \forall \alpha, s. BK(s) \wedge BK(do(\alpha, s))$*

The first assumption is the belief equivalent to Reiter's executable situation [Reiter, 1991]. Thus we assume that an action is only executed if the precondition is believed. This thus not restricts the possibility that later on the agent realize that the precondition of this action was not fulfilled.

The second assumption just state that the invariant definitions and the definition of the action effects are not contradicting. This is necessary as otherwise the agent could create inconsistency just by execution of actions without considering the world. This assumption was also used in [Steinbauer and Mühlbacher, 2014] in order to state theoretical results about history based diagnosis.

As a theoretical result we can state Lemma 1.

Lemma 1. *As long as the following does not hold if an action returns with true as result $\mathcal{D}^{cons}(\{\}) \models \neg SF(\alpha, s)$ and if an action returns with false as result $\mathcal{D}^{cons}(\{\}) \models SF(\alpha, s)$ it follows that the cardinality of the preferred diagnosis is 0.*

Proof (Sketch). A direct result of the definition of $Cons$ and the fact that the sensing is not contradicting with the situation. □

This lemma states that as long as no sensing contradicts the current belief the situation remains unchanged. This is of especially interest as it shows that without any indication that a fault has happened an agent using the consistency history based diagnosis will draw the same conclusions as if the agent would use the standard situation calculus. Thus only in case of an evidence of a fault the agent revises its belief.

We can give a simple upper bound on the minim number of faulty actions formally stated in Lemma 2.

Lemma 2. *The minimum number of faults has as upper bound the number of sensing actions within the history.*

Proof (Sketch). This is a direct result of Assumption 2 and the definition of $Cons$. □

Through Lemma 2 we can conclude Lemma 3.

Lemma 3. *The cardinality of the situations which are preferred diagnosis are bound by $\binom{|s|}{\#\alpha \in s \wedge \alpha \in \mathcal{A}_{sensing}}$.*

Proof (Sketch). This is a direct result of Lemma 2 □

This lemma shows that the number of sensing actions have a direct impact of the complexity of the approach. The more sensing actions are within the history the higher the complexity. It is important to mention that this upper bound can only be reached if all sensing action failed within the history. This can be stated more restricted in the following Theorem 1.

Theorem 1. *The cardinality of the situations which are minimal diagnosis are bound by $\binom{|s|}{\#faulty\ sensing\ actions}$. Where faulty sensing actions(α) is defined as: if an action returns with true as result and $Belief(\neg SF(\alpha), s)$ holds or if an action returns with false as result and $Belief(SF(\alpha), s)$.*

Proof (Sketch). Proof by induction

For the base case we assume no fault has happened thus through Lemma 1 the theorem holds.

For the induction step. Let's assume we have a history with n faulty sensing actions. The action α is performed and the sensing results contradict with the belief of the agent. Thus it follows that the minimal extension of any preferred diagnosis is due to a failing α. Due to the fact that we search for minimal diagnosis this bound must also hold for all new diagnosis thus the theorem follows. □

This theorem shows that the complexity of this approach is basically influenced by the number of faults the agents observes and the length of the history the agent considers being relevant.

In order to calculate the diagnosis one could simple try every action. A more advanced method is to use the conflict set to focus the search on those actions which may explain the discrepancy [Reiter, 1987].

4 Calculation of minimal diagnosis using conflicts

To efficiently find all diagnosis one want to avoid enumerating all possible sets of faulty actions and afterwards checking for each of these sets the consistency. In fact one can reverse the approach and use the information of the consistency checks to guide the search. This idea is also the basic principle behind the consistency based diagnosis as it was proposed by Reiter[Reiter, 1987].

As we use the same assumption as in [Steinbauer and Mühlbacher, 2014] it holds that a contradiction is always cased through a sensing action. Furthermore we are interested in actions which are faulty. Thus the conflict sets we are interested are sets of actions which cannot occur together within a history. In order to calculate these conflict sets we will first show how to calculate the conflict set for a history with one sensing action which is in contradiction. Afterwards we will show how this approach can be generalized to any history.

Let's assume we have a history $[\alpha_1, \alpha_2, \alpha_1^S, \alpha_3]$ were $\alpha_1, \alpha_2, \alpha_3$ are primitive actions and α_1^S is a sensing action. Furthermore let's assume we have performed an additional sensing action α_2^S which is in contradiction to the history. As a contradiction is always caused by at least one conflict of two formulas within the history we examine these conflicts. As sensing action assert the formula SF to the theory, and the theory was consistent before it follows that this formula must be one part of these conflicts. The second part is either a successor state axiom, another sensing axiom or an axiom of the initial situation as those determines the truth value of the fluent.

Let's consider the first possibility for a conflict the successor state axiom. As this axiom is guarded through an assumption of not faulty actions we can used these to create our conflict set. To derive the conflict first it needs to be checked if $\psi_\alpha^F(\bar{x}, s)$ holds. If this holds the action influence the fluent and all the actions yielding to the conclusion that the precondition of the action holds are part of the conflict set. If one of these actions would be assigned faulty the precondition may not hold and the fluent may get an arbitrary value. Additionally to these actions α itself is part of the conflict as a fault of this action may result in an arbitrary value for the fluent. Last but not least the sensing action causing the conflict is also part of the conflict.

If we consider the second possibility for a conflict we have a conflict between the sensing axiom which should be asserted and an already asserted sensing axiom. Thus two sensing axioms contradict each other thus the two actions are a conflict set. Additionally all the actions providing the precondition of the sensing action are part of the conflict set.

Finally let's consider the third possibility for a conflict which is the initial situation. As we assume that the initial situation is no faulty the only explanation is that the sensing action is wrong. Thus the sensing action is part of the conflict combined with the actions which support the precondition of the sensing action.

As the conflict is not only with the last action in the history the conflicting sensing formula SF is regressed back and the same conflict set search is applied.

In order to calculate all diagnosis Algorithm 1 can be applied to find a new set of conflict sets for a given diagnosis. In order to calculate these conflict sets Algorithm 2 is used which performs the conflict calculation outlined above.

Algorithm 1: ComputeAllConflictSets

```
Data: s ... a history
Data: δ ... a diagnosis to consider
Result: C ... a set of conflict sets
begin
    C = {} ;
    for s' ← do(α, s_0) to s do
        s' = do(α, s'') ;
        if α ∈ A_Sensing then
            if D^cons(δ) ∪ SF(α, s'') ⊨ ⊥ then
                C = ComputeConflictSet(s', α, SF(α, s''), δ) ∪
                    ComputeConflictSet(s', α, ¬Π_α, δ) ;
                break ;
            end
        end
    end
end
```

Algorithm 1 performs a search starting from the situation after performing the first action till the end of the history. In each iteration the algorithm checks if the action is a sensing action and if this action leads to a contradiction. If a contradiction is caused through this action the conflict set is calculated through algorithm 2.

Algorithm 2: ComputeConflictSet

```
Data: s ... a history
Data: α ... the action imposing φ
Result: φ ... an inconsistent formula
Data: δ ... a diagnosis to consider
begin
    C = {α} ;
    for s' ← s to s_0 do
        if s' = s_0 then
            break ;
        end
        ⟨r, c⟩ = checkSat(D^cons(δ) ∪ φ) ;
        if r = sat then
            break ;
        end
        s' = do(α', s'') ;
        if successorStatAxiomInConflict(c, α') then
            C = C ∪ ComputeConflictSet(s'', α', ¬Π'_α, δ)
        end
        if sensingAxiomInConflict(c, α') then
            C = C ∪ ComputeConflictSet(s'', α', ¬Π'_α, δ)
        end
        φ = Regress(φ, α')
    end
end
```

Algorithm 2 searches from the end of the history towards the initial situation. In each iteration the algorithm checks if the theory is consistent. If the theory is consistent no further part of the conflict can be derived. If a conflict exists it is checked if the conflict contains a successor state axiom which is influenced through the current action α'. If the con-

flict contains such an influence all the actions are searched which support the precondition of the action. If a conflict contains the sensing axiom of the current action all actions supporting the preconditions are searched again for α'.

In order to use the calculated conflicts set we need to show that the algorithm finds at least a supper set of each minimal conflict set. We state this property formally in the following theorem.

Theorem 2. *For every minimal conflict set C the Algorithm 1 finds a conflict set C' such that $C \subseteq C'$.*

Proof (Sketch). Poof by contradiction. Let's assume there exists a minimal conflict set C where the algorithm produces no correct C'.

As shown in [Steinbauer and Mühlbacher, 2014] only sensing action can indicate a fault. Thus a sensing action needs to be part of C. Furthermore as after a sensing produces an inconsistent theory and the definition of a minimal conflict only the first sensing action leading to an inconsistency is of interest. Thus it is sufficient only to consider s' and adding the sensing action to the conflict set.

By definition of the consistency a sensing action can only lead to an inconsistency iff the precondition holds for the sensing action and the negation of the sensing action is part of the situation (s') in which the action was performed. The precondition of an action holds in a situation iff the negation of the precondition yields add to the theory a contradiction. Thus what remains to show is that there exists an action which is not part of the result of Algorithm 2 for certain formula ϕ.

We show this by induction. For the base case the contradiction holds by the definition of the consistency, as it assumes that the initial situation is not contradicting. Thus no further actions can be in the conflict set.

The inductive step let's assume for every formula ϕ with any situation s' up to the length N the contradiction holds. Furthermore by the definition of regression Line 18 ensures that the formula ϕ is correctly transformed closer to the initial situation. Thus any conflicting action is found for s. Thus what remains to show is that the action α, such that $s = do(\alpha, s')$ holds, is the missing one in the conflict set. By definition of the CBAS it holds that an action has only an influence that a formula holds if it either is an effect of the action or if it is asserted through a sensing axiom. For the first case the action would have been found as the conflict would have been with a successor state axiom. For the second case the action would have been found as the conflict would have been with the sensing axiom. Thus there is no action missing in the conflict set. Thus the theorem follows. □

With the help of the conflict set calculation one can now apply a minimal hitting set algorithm to find all the diagnosis. Thus the calculation reduces to a standard diagnosis problem.

As the agent is interested in the belief and not only in the diagnosis one needs to perform a two step approach. The first step is to calculate the diagnosis and the second step is to use these diagnosis to reason about the belief. To avoid this two step approach a non monotonic reasoning needs to be applied. This can be done with the help of default logic. As default logic [Reiter, 1980] is known to be a method to calculate a diagnosis [Reiter, 1987].

5 Default logic to reason about the belief

The default logic consists of two parts. A set of logical axioms and a set of default rules. In order to use the default logic to calculate the diagnosis we use default rules. These default rules state that an action is not faulty if it does not cause a contradiction. The default rule θ is defined as follows:

$$\frac{: \neg fault(s)}{\neg fault(s)} \tag{2}$$

Additionally to the default rules we need a cardinality constraint Γ in order to only use the minimal diagnosis. Through Theorem 1 the exact cardinality is given through the number of contradicting sensing.

Finally the default logic theory $\mathcal{D}^D$ is defined as the pair $\langle \mathcal{D}_{ap} \cup \mathcal{D}_{ssa'} \cup \mathcal{D}_{una} \cup \mathcal{D}_{S_0} \cup \mathcal{D}_{AO} \cup \Sigma \cup Cons \cup \Gamma; \theta \rangle$ where $Cons$ is the consistency definition quantified over all situations. For the reasoning with the default logic one uses so called extensions [Reiter, 1980]. The following corolla can be stated for the extensions of the default logic.

Corolla 1. *Every maximal extension E of $\mathcal{D}^D$ correspond to a minimal diagnosis δ where $\delta = \{s | fault(s) \in E\}$*

Proof (Sketch). Through the definition of default logic is ever maximal extension a subset minimal set of faulty action. Through the cardinality constrain only cardinality minimal diagnosis are allowed. Furthermore an extension needs to be consistent through the definition of $Cons$ which is part of the theory. □

To answer a query for the belief on can now define $Belief^D(\phi, s) \equiv \mathcal{D}^D \models_{cautious} \phi(s)$. Thus a formula is belief if it holds in every extension of the default logic. We can state that the two definitions for the belief are equivalent in the following lemma.

Lemma 4. *A formula ϕ is belief through $Belief^D(\phi, s)$ iff $Belief(\phi, s)$ holds for the formula.*

Proof (Sketch). This directly follows from Corolla 1 and the reason that we perform a cautious reasoning. □

6 Relation to fault based belief management

As we argued in the introduction it is not always feasible to describe how an action fails. But what happens if we can model the faults. Would this approach yield the same results as if we would use a fault model based one as for example described in [Gspandl *et al.*, 2011]?

In order to answer this question we first show how to define the influence of an action if we have the knowledge about fault modes. If we have given for an action α the fault modes α_{F_1} and α_{F_2} we create the formula ψ^F_α as follows:

$$\begin{aligned}\psi^F_\alpha(\bar{x}, s) \equiv & \exists \bar{y}.\phi^+_F(\bar{x}, \alpha_{F_1}(\bar{y}), s) \vee \exists \bar{y}.\phi^-_F(\bar{x}, \alpha_{F_1}(\bar{y}), s) \vee \\ & \exists \bar{y}.\phi^+_F(\bar{x}, \alpha_{F_2}(\bar{y}), s) \vee \exists \bar{y}.\phi^-_F(\bar{x}, \alpha_{F_2}(\bar{y}), s)\end{aligned} \tag{3}$$

Where ϕ_F is the formula corresponding to fluent F. Thus the influenced fluents are those which would be influence through a fault event by any possible assignment of variables to the actions. Through this construction the following lemma follows.

Lemma 5. *Given a situation s and a fault mode assignments of action A_F, which makes the situation consistent then it follows that $\mathcal{D}^{cons}(\delta) \models \phi(s) \Rightarrow \mathcal{D} \models \phi(s')$, where δ consist of those actions which have a fault mode assigned and s' is the result of replacing all action in s through their fault mode.*

Proof (Sketch). Proof by contradiction. Let's assume there exists a ϕ which is entailed by $\mathcal{D}^{cons}(\delta)$ but not by $\mathcal{D}$ then it follows that there exists at least on fluent F which has a truth assignment which differs in both reasoning systems. As the reasoning is the same for those action which are not faulty it follows that there exists an action α, which is fault, which sets the truth value of the fluent differently. As α is fault no truth value is set through the definition of the successor state axiom. Thus such a ϕ can't exists. □

The opposite direction does not hold thus the method describe in this paper is more cautions in answering a query. As the reasoning the agent uses is based on the belief we need to consider how the belief is created. The belief is based on reasoning over all minimal diagnosis. Thus we need to relate the diagnosis to state how the belief of an agent with fault modes and without fault modes is related. The following formal lemma holds for the diagnosis of the different belief managements.

Lemma 6. *Given a situation s and a fault mode assignments of action A_F, which makes the situation consistent then it follows that δ consisting of all actions which have a fault mode assigned is a diagnosis.*

Proof (Sketch). This directly follows from Lemma 5. And the definition of a diagnosis is a set of action which blamed to be fault create a consistent situation. □

As in the case above the opposite direction does not hold in general thus the method described in this paper accepts diagnosis which would not be consistent if fault modes would be used. This also yield the result that there is not connection of minimal diagnosis between the two methods. This can be shown be the following simple example.

Let's assume we have an action α which triggers that fluent F_a and F_B is true afterwards. Furthermore let's assume we have two sensing actions α_a^S and α_b^S. Action α_a^S sense fluent F_a and action α_b^S fluent F_B. The fault modes are that both sensing action sense the fluent value wrongly. Additionally the action α can fail with α_{F_1} and α_{F_2}. α_{F_1} sets fluent F_a to false and F_B to true. α_{F_2} sets fluent F_a to true and F_B to false.

If we would now observe the following action sequence $[\langle\alpha, \top\rangle; \langle\alpha_a^S, \bot\rangle]$, we have as minimal diagnosis $\{\alpha_{F_1}\}$ and $\{\alpha_a^S\}$. This diagnosis would also be minimal in case if no fault modes would be used. We can extend the sequence by executing α_a^S which would for example $[\langle\alpha, \top\rangle; \langle\alpha_a^S, \bot\rangle; \langle\alpha_b^S, \bot\rangle]$. This would cause to calculate new minimal diagnosis as none of the above one is still consistent and would yield the only minimal diagnosis $\{\alpha_a^S, \alpha_b^S\}$. On the contrary if no fault modes would be used blaming action α_{F_1} would be sufficient as it influence F_a and F_b.

As the minimal diagnosis have no connection to each other the belief of an agent can diverge depending if fault modes are used or not. In the example above with the action sequence $[\langle\alpha, \top\rangle; \langle\alpha_a^S, \bot\rangle; \langle\alpha_b^S, \bot\rangle]$ the agent would belief F_a and F_b if fault modes would be used. If no fault modes would be used the agent would belief $\neg F_b$ as α_b^S would not be blamed to be faulty.

7 Related Research

Before concluding the paper we want to discuss related and influencing work.

Our work is based on the idea of altering the history of an agent to create a consistent history. The idea as it was published in [Delgrande and Levesque, 2012] and [Gspandl *et al.*, 2011] use defined variations of actions to change the history. Thus a precise diagnosis could be calculated with the need of a precise model what can go wrong. Additionally it should be considered that in contrast to our approach every instantiation of a variation needs to be considered which can lead to an infinite space of diagnosis. In contrast to our approach this can only yield an exponential number of diagnoses. Even in case of a finite number of variations our method has a lower complexity due to the two states an action can have. This is in contrast to many different states in case of fault modes.

The main influence for our work is based on consistency based diagnosis as it was described in [Reiter, 1987]. The idea is to blame component to be faulty and to assume any behavior of faulty components. This idea was extended in [Mcllraith, 1999] and [Baral *et al.*, 2000] for dynamic system. The system can perform action and can make observations. Specific actions can make components fault. Thus a diagnosis is as set of actions making component faulty. In contrast to our approach this method searches for actions which cause a component fault, where as our method searches for failed actions. Furthermore the idea of consistency based diagnosis was translated in [Baier *et al.*, 2014] to use a possible world semantic. This translation allows planning for sensing action to discriminate a diagnosis or to plan a repair.

A work inspired by consistency based diagnosis and applying to execute actions was published in [Roos and Witteveen, 2005]. Within this work an executed plan together with partial observations about the world was diagnosed. To create a consistent view about the world actions where blamed to be abnormal as a diagnosis. The outcome of an action which is blamed to be abnormal was undefined. In contrast to our approach only those fluents are undefined which are set to a value in case of the action was executed successfully. Thus it is not possible to define which influence a failing action could have. Thus one could not model that a faulty block stacking can change the positions of the block in hand and other blocks on the table. The approach was further extended in [Witteveen *et al.*, 2005] to calculate the diagnosis in a distributed way in a multiagent system. Additionally the root cause of the failing action could be more discriminated through the diagnosis of components (the environment or an agent, ...) which were involved in faulty actions. An important difference to our approach is that we use the diagnosis to model the belief of the agent for further reasoning.

A simply way of dealing with action outcomes which are not used any more was presented in [Rajaratnam *et al.*, 2014]. The authors show how the effect of a sensing action can be forgotten. In contrast to our approach only sensing action can be forgotten and not the effect of any other action.

As discussed above the idea of guarding the successor state axioms in not new. In some definitions of the succes-

sor state axioms [Reiter, 1991] the action precondition was used as a guard. This work was extended in [Sardina, 2000] [De Giacomo *et al.*, 2001] to evaluated sensing an action outcomes with guarded conditions. We took this idea to narrow the possible influence of a faulty action outcome. Thus extending the previous work with the diagnosis process of faulty actions.

Our approach uses a possible world semantic combined with a successor state axiom which allows partial knowledge. This method allows expressing many different partial known states. But it comes with the costs of possible world reasoning. In [Petrick, 2008] the authors proposed the use of Cartesian situation to define knowledge. In contrast to possible worlds each fluent can have an undefined states. All possible combination of all fluents can be considered as valid combinations. Through this compact representation a realization with the help of a Strips like table approach as it was described in [Petrick and Bacchus, 2002] can be used. This formalism is powerful enough to be used for challenging planning problems. The main disadvantage of this approach is that complex constraints imposed by sensing, as it is used in our work, cannot be used. Thus a sensing can only define the value of fluent not a constrain between values of two fluents.

8 Conclusion and Future Work

Within this paper we showed how to formalize a belief management for an agent which uses the history of performed action. Furthermore the formalization does not require any specification of known faults for an action. Thus the approach can be simple applied to agents which use a belief based reasoning.

An important future step is to evaluate the approach if the reasoning is tractable for real world applications. Furthermore it would be of interest if fault modes can be used in addition for those action faults which are known. Thus one could add as much information for the belief management as currently available about action effects.

References

[Baier *et al.*, 2014] J Baier, Brent Mombourquette, and Sheila A McIlraith. Diagnostic problem solving via planning with ontic and epistemic goals. In *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 2014.

[Baral *et al.*, 2000] Chitta Baral, Sheila McIlraith, and Tran Cao Son. Formulating diagnostic problem solving using an action language with narratives and sensing. In *KR*, pages 311–322, 2000.

[De Giacomo *et al.*, 2001] Giuseppe De Giacomo, Hector J Levesque, and Sebastian Sardina. Incremental execution of guarded theories. *ACM Transactions on Computational Logic (TOCL)*, 2(4):495–525, 2001.

[Delgrande and Levesque, 2012] James P Delgrande and Hector J Levesque. Belief revision with sensing and fallible actions. In *KR*. Citeseer, 2012.

[Gspandl *et al.*, 2011] Stephan Gspandl, Ingo Pill, Michael Reip, Gerald Steinbauer, and Alexander Ferrein. Belief management for high-level robot programs. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 900, 2011.

[McCarthy, 1963] J. McCarthy. Situations, Actions and Causal Laws. Technical report, Stanford University, 1963.

[Mcllraith, 1999] Sheila A Mcllraith. Explanatory diagnosis: Conjecturing actions to explain observations. In *Logical Foundations for Cognitive Agents*, pages 155–172. Springer, 1999.

[Petrick and Bacchus, 2002] Ronald PA Petrick and Fahiem Bacchus. A knowledge-based approach to planning with incomplete information and sensing. In *AIPS*, pages 212–222, 2002.

[Petrick, 2008] Ronald PA Petrick. Cartesian situations and knowledge decomposition in the situation calculus. *KR*, 8:629–639, 2008.

[Rajaratnam *et al.*, 2014] David Rajaratnam, Hector J Levesque, Maurice Pagnucco, and Michael Thielscher. Forgetting in action. *Proc. KR*, 2014.

[Reiter, 1980] Raymond Reiter. A logic for default reasoning. *Artificial intelligence*, 13(1):81–132, 1980.

[Reiter, 1987] Raymond Reiter. A theory of diagnosis from first principles. *Artificial intelligence*, 32(1):57–95, 1987.

[Reiter, 1991] Raymond Reiter. The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. *Artificial intelligence and mathematical theory of computation: papers in honor of John McCarthy*, 27:359–380, 1991.

[Reiter, 2001] R. Reiter. *Knowledge in Action. Logical Foundations for Specifying and Implementing Dynamical Systems*. MIT Press, 2001.

[Roos and Witteveen, 2005] Nico Roos and Cees Witteveen. Diagnosis of plans and agents. In *Multi-Agent Systems and Applications IV*, pages 357–366. Springer, 2005.

[Sardina, 2000] Sebastian Sardina. *Indigolog: Execution of guarded action theories*. PhD thesis, Citeseer, 2000.

[Scherl and Levesque, 1993] Richard B Scherl and Hector J Levesque. The frame problem and knowledge-producing actions. In *AAAI*, volume 93, pages 689–695. Citeseer, 1993.

[Steinbauer and Mühlbacher, 2014] Gerald Steinbauer and Clemens Mühlbacher. Using common sense invariants in belief management for autonomous agents. In *2014 AAAI Spring Symposium Series*, 2014.

[Witteveen *et al.*, 2005] Cees Witteveen, Nico Roos, Roman van der Krogt, and Mathijs de Weerdt. Diagnosis of single and multi-agent plans. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 805–812. ACM, 2005.