Temporal Projection and Explanation

Andrew B. Baker and Matthew L. Ginsberg*
Computer Science Department
Stanford University
Stanford, California 94305

Abstract

We propose a solution to problems involving temporal projection and explanation (e.g., the Yale shooting problem) based on the idea that whether a situation is abnormal should not depend upon historical information about how the situation arose. We apply these ideas both to the Yale shooting scenario and to a blocks world domain that needs to address the qualification problem.

1 Introduction

The paper [1987] by Hanks and McDermott describing the Yale shooting problem has generated such a flurry of responses that it is difficult to imagine what another one can contribute. The points raised by Hanks and McDermott, both formal and philosophical, have been discussed at substantial length elsewhere.

It is not our intention to enter here into the philosophical portion of this discussion. Rather, the contribution we hope to make is technical. In [], the authors raise a specific technical question: How is a certain form of temporal reasoning to be modelled? They go on to demonstrate that none of the then existing descriptions of action using nonmonotonic reasoning is suitable, and proceed from this observation to draw more far-reaching conclusions regarding the role of formal methods in AI. A discussion of these philosophical issues is outside the intended scope of this paper.

Since the initial circulation of Hanks' and McDermott's ideas in 1985, a variety of researchers have proposed solutions to the technical challenge they raised [, , , , , , , , and others]. Unfortunately, subsequent analysis (most of it over coffee tables) has shown most of these intended solutions to be flawed for one or another technical reason.

We begin in the next section by presenting a brief description of the Yale shooting problem itself, and a variant that appears in []. The description given is in fact somewhat simpler than that presented in [], but still retains all of the troublesome features of the original. In Section 3, we go on to describe proposed solutions, and investigate their technical shortcomings.

The formal underpinnings of our own ideas are presented in Section 4, and we return to the Yale shooting scenario in Section 5, showing that our notions can be used to solve both the original problem and the variant presented in Section 2. In Section 6, we extend our ideas to deal with the qualification problem in a simple blocks world scenario. Concluding remarks are contained in Section 7.

*This work has been supported by the Rockwell Palo Alto Laboratory, by General Dynamics, and by NSF under grant number DCR-8620059. The first author is supported by an AFOSR graduate fellowship.

2 The Yale shooting

The Yale shooting problem involves reasoning about a sequence of actions. In order to keep our notation as manageable as possible, we will denote the fact that some fluent f holds in a situation s by writing simply

$$\mathtt{holds}(f, s).$$

A situation will be written as an ordered sequence of actions

$$s = \langle a_n, \ldots, a_1 \rangle.$$

(We write the last action first so that we can append new actions to the front of the sequence.) The initial situation is written simply as $\langle\rangle$, corresponding to the empty sequence of actions.

The Hanks-McDermott problem involves an individual, often named Fred, and a gun. Fred may or may not be alive, and the gun may or may not be loaded. In the initial situation, however, Fred is alive and the gun is loaded:

$$\mathtt{holds}(\mathtt{alive}, \langle\rangle) \tag{1}$$

$$\mathtt{holds}(\mathtt{loaded}, \langle\rangle) \tag{2}$$

We are also told that in any situation in which the gun is loaded, firing it will kill Fred. We write this as:

$$\mathtt{holds}(\mathtt{loaded}, s) \supset \neg\mathtt{holds}(\mathtt{alive}, \langle \mathtt{shoot} | s \rangle), \tag{3}$$

where we are using $\langle a|s \rangle$ to construct the result of performing the action a in the situation s. Finally, we are given a frame axiom:

$$\neg\mathbf{ab}(a, f, s) \supset [\mathtt{holds}(f, s) \equiv \mathtt{holds}(f, \langle a|s \rangle)]. \tag{4}$$

In other words, unless the action a is abnormal in that it potentially reverses the fluent f in the situation s, f will persist through the execution of the action. We will presumably minimize the extent of the abnormality predicate using any of the formal approaches to nonmonotonic reasoning [].[1]

Now suppose that in the initial situation, we wait and then fire the gun. Does Fred die? In other words, what can we say about

$$\texttt{holds}(\texttt{alive}, \langle \texttt{shoot}, \texttt{wait} \rangle)?$$

Remarkably, we can say very little about it. Our *intuition* is that the shooting action should be abnormal, and should cause Fred's aliveness to be terminated. In other words, we should be able to derive

$$\texttt{ab}(\texttt{shoot}, \texttt{alive}, \langle \texttt{wait} \rangle)$$

from the domain description given by (1)–(4).

Unfortunately, we cannot do this. As Hanks and McDermott noticed, all that we can actually derive is the disjunction

$$\texttt{ab}(\texttt{shoot}, \texttt{alive}, \langle \texttt{wait} \rangle) \vee \texttt{ab}(\texttt{wait}, \texttt{loaded}, \langle \rangle). \quad (5)$$

This is saying that *either* Fred becomes not alive after the shooting action, *or* the waiting action causes the gun to become unloaded. If our intention is merely to minimize abnormality, there is no way to select between these two possibilities.[2] As this sort of a minimization is all that is effected by any of the existing formalizations of nonmonotonic reasoning, all of these formalizations are incapable of concluding that Fred is dead after we shoot him.

Causal reasoning of this sort, where one reasons forward in time from a fixed initial state, is called "temporal projection" by Hanks and McDermott in [1987]. Before we turn to the various solutions that have been proposed to the temporal projection problem, let us discuss its dual, the temporal *explanation* problem [].

This involves reasoning backward from information given about some state in the future. Here is an example similar to one introduced by Kautz in [1986]:

In this example, which we call the "murder mystery," we drop the axiom (2) saying that the gun is loaded in the initial situation, and replace it with a statement to the effect that if we fire the gun and *then* wait, Fred will be dead:

$$\neg\texttt{holds}(\texttt{alive}, \langle \texttt{wait}, \texttt{shoot} \rangle).$$

Given this information, is it legitimate to conclude that the gun was loaded in the initial situation after all?

[1] Our formulation is slightly different from that presented in []. We use it because: (1) It is as natural a description of the domain as that given by Hanks and McDermott in [1987], (2) In the original description, it was impossible to conclude $\neg\texttt{holds}(\texttt{alive}, \langle \texttt{shoot}, \texttt{wait} \rangle)$ in *either* extension, since there were no axioms relating alive to the fluent dead used in [], and (3) The description appearing in [] requires the introduction of an axiom explicitly stating that the shoot action leads to an abnormality for the alive predicate. This explicit analog to STRIPS delete lists will make it difficult to extend this approach to more complex domains [,].

[2] In our formulation, there is also a third possibility, namely $\texttt{ab}(\texttt{wait}, \texttt{alive}, \langle \rangle)$, where Fred dies while we wait to shoot him. This possibility will have no effect on the arguments to be presented in this paper.

One would expect so, although this conclusion, as in the original Yale shooting problem, is not sanctioned by circumscription. The reason is that we are unable to distinguish between extensions with

$$\texttt{ab}(\texttt{shoot}, \texttt{alive}, \langle \rangle) \quad (6)$$

where Fred dies when we shoot him, and those with

$$\texttt{ab}(\texttt{wait}, \texttt{alive}, \langle \texttt{shoot} \rangle),$$

where Fred dies acausally during the waiting action. Indeed, even if we could select the abnormality appearing in (6), it would not follow from this that the gun was actually loaded before the shooting action took place.

3 Solutions

The collection of proposed solutions to the original Yale shooting problem fall into roughly two groups. One set of researchers [, ,] suggests resolving the ambiguity underlying the problem by appealing to temporal information; another [, ,] suggests modifying the domain description so as to capture the notion of causality directly.

3.1 Temporal orders

The first idea involves selecting among competing extensions based on a temporal order. Thus we might prefer the extension in which $\texttt{ab}(\texttt{shoot}, \texttt{alive}, \langle \texttt{wait} \rangle)$ holds over that in which $\texttt{ab}(\texttt{wait}, \texttt{loaded}, \langle \rangle)$ holds on the grounds that $\texttt{ab}(\texttt{wait}, \texttt{loaded}, \langle \rangle)$ refers to an earlier time than does $\texttt{ab}(\texttt{shoot}, \texttt{alive}, \langle \texttt{wait} \rangle)$. This is the older of the two approaches, and is in fact endorsed by Hanks and McDermott as a possible solution in [1987]; the proponents of this approach argue that there can be no causal reason for the earlier abnormality.

Unfortunately, this approach can be used for temporal projection problems only. In the murder mystery, where we need to select between

$$\texttt{ab}(\texttt{shoot}, \texttt{alive}, \langle \rangle)$$

and

$$\texttt{ab}(\texttt{wait}, \texttt{alive}, \langle \texttt{shoot} \rangle),$$

the chronological approaches favor the second extension, a choice that is at odds with our intuition.

One might hope that one could use "reverse chronological minimization" to deal with temporal explanation problems, now preferring those extensions in which the abnormality occurs as *early* as possible. Unfortunately, reverse chronological minimization remains unable to conclude that the gun was actually loaded in the murder mystery. In addition, it seems extremely unlikely that problems of practical interest will admit a clear distinction between these two aspects of temporal reasoning.

3.2 Causal approaches

Other proposed solutions [, ,] have involved reformulating the original domain description in a way that captures the notion of causality directly. Loosely speaking, these approaches proceed by indicating *explicitly* that the shooting action causes Fred to become not alive, and by saying that no change can occur unless there is a causal explanation for that change.

Hanks and McDermott raise a variety of philosophical objections to this reformulation. Principally, they argue that nonmonotonic formalisms were introduced to enable us to give intuitive descriptions of things like the frame axiom. It would be unfortunate if they needed to resort to descriptions of causality that were (at least to Hanks' and McDermott's eyes) anything but intuitive after all.

In addition, it will in general be impossible to explicitly list all possible causal connections; some will be consequences of constraints on the situation that arises when an action is executed.[3] The causal approaches suggest that this problem can be addressed by minimizing the extent of the causes predicate instead of minimizing abnormality.

This does not work in all situations, however. The simplest one in which it fails is probably Hanks' and McDermott's original description of the Yale shooting problem, in which the gun is initially unloaded and there is a load action that loads it:

$$\mathtt{causes}(\mathtt{load}, \mathtt{loaded}, \mathtt{true}).$$

The intent of the above axiom is that the load action causes the fluent loaded to become true.

Now we also introduce a fluent dead, along with a domain constraint saying that Fred is dead if and only if he is not alive:

$$\mathtt{holds}(\mathtt{dead}, s) \equiv \neg\mathtt{holds}(\mathtt{alive}, s). \tag{7}$$

Finally, we have an axiom saying that shooting the gun causes Fred to become not alive:

$$\mathtt{causes}(\mathtt{shoot}, \mathtt{alive}, \mathtt{false}).$$

Now, we would like to be able to conclude that since shooting the gun causes Fred to become not alive, it also causes him to die:

$$\mathtt{causes}(\mathtt{shoot}, \mathtt{dead}, \mathtt{true}) \tag{8}$$

Unfortunately, this does not follow. Another possible minimization of the causes predicate is one in which the load action causes Fred to die, and the theory needs to be modified in a way that disallows this choice. Although a variety of suggestions have been made to address this problem, none of them is completely satisfactory at this point.

4 Situations and states

Note, however, that the causal approaches *do* correctly conclude that the gun was loaded in the murder mystery, since this is the only explanation that avoids the introduction of a new causal rule.

The causal approaches work because they describe actions not directly in terms of situations, but in terms of the values of various fluents in those situations. Because of this, they have no reason to introduce an abnormality into s in order to avoid one in $\langle \mathtt{wait}|s\rangle$.

In [], this observation led to the argument that the simplest way to solve the Yale shooting problem is simply to allow the function that computes the result of an action to vary. It is shown there that this idea can indeed be used successfully; the ideas we are about to present are a natural outgrowth of this earlier work.[4]

Another way to understand the problem is as follows: When we attempt to minimize abnormalities of the form

$$\mathbf{ab}_s(a, f, \langle a_n|\ldots|a_1\rangle),$$

we are not minimizing over situations so much as we are over action sequences; we might just as well be minimizing a function

$$\mathbf{ab}'_s(a, f, a_n, \ldots, a_1). \tag{9}$$

This sort of a minimization, however, completely misses the intuition that we are hoping to capture – our intent is to think of a situation not as a sequence of actions, but instead in terms of the values taken by various fluents.

In order to make this explicit, we will make a distinction between a *situation*, such as $\langle \mathtt{wait}\rangle$, and a *state*, which will be described simply in terms of fluent values. If a fluent f changes value when an action a is performed in a situation s, we will think of the abnormality not as a function of the situation s, but as a function of the state corresponding to it.

In order to formalize this, we introduce a subscript onto the frame axiom (4):

$$\neg\mathbf{ab}_s(a, f, s) \supset [\mathtt{holds}(f, s) \equiv \mathtt{holds}(f, \langle a|s\rangle)]. \tag{10}$$

The subscript indicates that it is a *situation* that is being considered abnormal.

In order to connect situations to the state vectors describing them, we introduce a predicate

$$\mathtt{describes}(s, v)$$

that is intended to capture the notion that the situation s is described by the state vector v. The intention here is that v contains enough information to determine the value assigned by s to any particular fluent f. In the Yale shooting domain, v might contain two entries, corresponding to the values of the alive and loaded fluents respectively; the value of any other fluent, such as dead, can be computed from these two.[5]

We now say that abnormal situations correspond to abnormal state vectors:

$$\mathbf{ab}_s(a, f, s) \wedge \mathtt{describes}(s, v) \supset \mathbf{ab}_v(a, f, v). \tag{11}$$

We also need an axiom saying that for any given state vector v, there is some situation $\mathtt{s}(v)$ that is described by it:

$$\mathtt{describes}(\mathtt{s}(v), v). \tag{12}$$

[3] This is known as the *ramification* problem [,].

[4] We also avoid several problems with the approach in []. Varying the value of a function is at odds with the unique names assumption, and makes it difficult to implement the suggestions in []; the work we are presenting assumes the validity of the unique names assumption, in accordance with most work in logic programming. We also avoid the introduction of the temporal order used in [], although this is a less substantial matter.

[5] The conditions that the state completely describe the situation and that the user supply the system with a description of the describes predicate are fairly restrictive ones. In a forthcoming paper [], one of us will discuss an extension of these ideas that uses partial descriptions only.

(Note that this axiom says only that *some* situation is described by the state vector v; in actuality many situations may be.)

Instead of minimizing $\mathtt{ab}_s$, we minimize $\mathtt{ab}_v$, allowing all other predicates to vary. That's all there is to it.

5 The shooting domain

As an example, we return to the Yale shooting domain, which we have axiomatized using (3), describing the shooting action, and (7), describing the predicate dead.

We also capture the describes predicate as:

$$\begin{aligned}&\mathtt{describes}(s,[\mathtt{alive},\mathtt{loaded}])\\&\quad\equiv \mathtt{holds}(\mathtt{alive},s)\wedge \mathtt{holds}(\mathtt{loaded},s)\\&\mathtt{describes}(s,[\mathtt{alive},\mathtt{unloaded}])\\&\quad\equiv \mathtt{holds}(\mathtt{alive},s)\wedge \neg\mathtt{holds}(\mathtt{loaded},s)\\&\mathtt{describes}(s,[\mathtt{dead},\mathtt{loaded}])\\&\quad\equiv \neg\mathtt{holds}(\mathtt{alive},s)\wedge \mathtt{holds}(\mathtt{loaded},s)\\&\mathtt{describes}(s,[\mathtt{dead},\mathtt{unloaded}])\\&\quad\equiv \neg\mathtt{holds}(\mathtt{alive},s)\wedge \neg\mathtt{holds}(\mathtt{loaded},s)\end{aligned}$$

We begin by noting that it is a consequence of the above axioms that

$$\mathtt{ab}_v(\mathtt{shoot},\mathtt{alive},[\mathtt{alive},\mathtt{loaded}]). \tag{13}$$

In other words, the shooting action is abnormal in any situation where Fred is alive and the gun is loaded.

To see this, suppose we denote by s_0 the situation $\mathtt{s}[\mathtt{alive},\mathtt{loaded}]$. It follows from (12) that

$$\mathtt{describes}(s_0,[\mathtt{alive},\mathtt{loaded}]),$$

so that

$$\mathtt{holds}(\mathtt{alive},s_0)\wedge \mathtt{holds}(\mathtt{loaded},s_0),$$

and therefore

$$\neg\mathtt{holds}(\mathtt{alive},\langle\mathtt{shoot}|s_0\rangle)$$

and

$$\mathtt{ab}_s(\mathtt{shoot},\mathtt{alive},s_0).$$

Applying (11), we obtain (13).

No other state vectors or actions will be abnormal if $\mathtt{ab}_v$ has its minimal extent, so we conclude that

$$\neg\mathtt{ab}_v(\mathtt{wait},\mathtt{loaded},[\mathtt{alive},\mathtt{loaded}]),$$

and therefore that

$$\neg\mathtt{ab}_s(\mathtt{wait},\mathtt{loaded},\langle\rangle),$$

where the initial situation $\langle\rangle$ is given by (1) and (2). It follows that the gun remains loaded during the wait action, and Fred dies. The murder mystery is also solved correctly, since it is only by assuming $\mathtt{holds}(\mathtt{loaded},\langle\rangle)$ that we can avoid increasing the extent of the $\mathtt{ab}_v$ predicate.

6 The blocks world

The ideas we have presented can be tested much more thoroughly if we consider a somewhat more complex domain of discourse. Let us therefore turn our attention to a simple blocks world domain containing three blocks, A, B and C.

In this domain, there is an action $\mathtt{move}(b,l)$ that takes a block b and relocates it at a location l. However, rather than treat the requirements that both the block and the target location be clear as preconditions to the move action, we will view them as qualifications to it. This leads us to describe the move action as:

$$\neg\mathtt{ab}_{qs}(\mathtt{move}(b,l),s)\supset \mathtt{holds}(\mathtt{loc}(b,l),\langle\mathtt{move}(b,l)|s\rangle),$$

where the $\mathtt{loc}(b,l)$ means that the block b is located at l.

The subscripts q and s on the abnormality predicate indicate respectively that it is a qualification abnormality that is being considered, and that the qualification is being described for situations as opposed to state vectors. As in Section 4, we introduce:

$$\mathtt{ab}_{qs}(a,s)\wedge \mathtt{describes}(s,v)\supset \mathtt{ab}_{qv}(a,v), \tag{14}$$

and minimize $\mathtt{ab}_{qv}$ instead of $\mathtt{ab}_{qs}$.

In order to keep our notation uniform, we introduce a subscript f into the "frame" axioms (10) and (11), obtaining

$$\neg\mathtt{ab}_{fs}(a,f,s)\supset[\mathtt{holds}(f,s)\equiv \mathtt{holds}(f,\langle a|s\rangle)] \tag{15}$$

$$\mathtt{ab}_{fs}(a,f,s)\wedge \mathtt{describes}(s,v)\supset \mathtt{ab}_{fv}(a,f,v) \tag{16}$$

Now when we do the minimization, we minimize $\mathtt{ab}_{qv}$ preferentially over $\mathtt{ab}_{fv}$. If we were not to do so, then we could avoid all violations of the frame axiom simply by assuming that all actions were qualified! We write our circumscription policy as

$$\mathtt{ab}_{qv} > \mathtt{ab}_{fv}. \tag{17}$$

Returning to the blocks world, we have the domain constraints

$$\mathtt{holds}(\mathtt{loc}(b,l),s)\wedge(b\neq b')\supset\neg\mathtt{holds}(\mathtt{loc}(b',l),s) \tag{18}$$

$$\mathtt{holds}(\mathtt{loc}(b,l),s)\wedge(l\neq l')\supset\neg\mathtt{holds}(\mathtt{loc}(b,l'),s) \tag{19}$$

$$\neg\mathtt{holds}(\mathtt{loc}(b,b),s) \tag{20}$$

saying respectively that only one block can be in any particular location, that a block can be in only one place at a time, and that no block can be on top of itself.

Next, we describe a situation by giving the locations of the three blocks:

$$\begin{aligned}\mathtt{describes}(s,[a,b,c])&\equiv \mathtt{holds}(\mathtt{loc}(\mathtt{A},a),s)\\&\wedge \mathtt{holds}(\mathtt{loc}(\mathtt{B},b),s)\wedge \mathtt{holds}(\mathtt{loc}(\mathtt{C},c),s).\end{aligned} \tag{21}$$

Finally, although the move action has no explicit preconditions, it can be qualified if either the block being moved or the target location is not clear:

$$\begin{aligned}\mathtt{holds}(\mathtt{loc}(x,b),s)&\supset \mathtt{ab}_{qs}(\mathtt{move}(b,l),s)\\\mathtt{holds}(\mathtt{loc}(x,l),s)&\supset \mathtt{ab}_{qs}(\mathtt{move}(b,l),s)\end{aligned}$$

At this point, our description is complete. (The action of moving a block to itself is qualified by virtue of the fact that the success of the action is in direct conflict with (20).) But there is one additional subtlety we need to consider. Consider again the axiom (12) saying that every description corresponds to some situation; what situation corresponds to the state $[l_1, l_1, l_2]$, where both `A` and `B` are located at l_1?

In light of the domain constraint (18) indicating that only one block can be at any particular location, there is no such situation. It follows that (18) and (12) are in conflict; what we really wanted to say in (12) is that every *consistent* description has some associated situation. We therefore weaken (12) to become

$$\neg\mathtt{inconsistent}(v) \supset \mathtt{describes}(\mathtt{s}(v), v) \tag{22}$$

and minimize `inconsistent`. Since whether or not a state vector is inconsistent should be determined before considering the qualification or frame problem, our revised circumscription policy is

$$\mathtt{inconsistent} > \mathtt{ab}_{qv} > \mathtt{ab}_{fv}. \tag{23}$$

Note that the predicate `inconsistent` appears nowhere else in our domain description; specifically, we do *not* need to axiomatize a notion of consistency. The determination of the consistency of any particular state vector is handled automatically by the circumscription.

Note also that the ramification problem is handled by this approach. Thus, for example, if we move a block from one location to another, we are able to conclude that the block is no longer at its original location *without* introducing an additional axiom to this effect – the conclusion is a consequence of the relocation, together with the axiom (19) saying that the block can be in only one location at a time.

7 Conclusion

We hope that the simplicity of the solution we have suggested will defuse some of the arguments that McDermott has recently raised against formal work on commonsense reasoning []. Indeed, the description of action we have developed seems rather *more* robust than a procedural description of action such as that proposed in []. The ramification problem is well-treated, we do not need to supply an explicit "delete" list of those facts that an action makes abnormal, and it is possible to incorporate qualifications into our description. To the best of our knowledge, the description we have presented is the *only* one capable of representing the blocks world of Section 6 using a nonmonotonic description of the frame axiom and an implicit description of ramifications in terms of domain constraints.

Much work remains to be done, however. For example, all of the situations we have examined in this paper have been "complete," in that we have been able to compute the state vectors associated with them. If we are to incorporate our description of action into a general-purpose planner, however, we will need to work with situations about which we have only partial knowledge, such as situations in which the goal has been achieved, or in which some subgoal has been accomplished. Preliminary results in this direction are the focus of [].

We will also need to extend our results to deal with more sophisticated ontologies of time and action. Although space does not allow us to discuss these matters here, these issues present no substantial difficulty for the approach that we have described. Essentially, the contribution we have made is to notice that we must avoid being syntactically trapped into minimizing abnormality over actions as in (9), and must instead be careful to ensure that we perform the minimization over the combinations of fluents and values to which these action sequences correspond. This observation is independent of the representation we choose for time or action, and extends easily to integer- or real-valued time and to concurrent actions.

Acknowledgement

We are grateful to Pat Hayes, David Israel, Paul Morris, Karen Myers, Eunok Paek, Yoav Shoham and Dave Smith for many stimulating discussions, and are especially indebted to Vladimir Lifschitz for careful conversations during a Stanford faculty retreat.

References

[]
[]
[]
[]
[]
[]
[]
[]
[]
[]
[]
[]
[]
[]