Explanation in the Situation Calculus

Murray Shanahan
Imperial College
Department of Computing,
180 Queen's Gate,
London SW7 2BZ.
England.

Abstract

This paper explores different techniques for explanation within the framework of the situation calculus, using the so-called stolen car problem as its main example. Two approaches to explanation are compared: the deductive approach usually found in the literature, and a less common abductive approach. Both approaches are studied in the context of two different styles of representation.

Introduction

A great deal of attention has been given to the problem of formalising prediction, that is reasoning forwards in time from causes to effects, and in particular to the logical aspect of the frame problem. Fewer authors, however, have studied the converse problem of formalising temporal explanation (or postdiction), that is reasoning backwards in time from effects to causes. Temporal explanation is certainly as important as prediction, as it underlies planning and diagnosis, as well as being a fundamental mode of reasoning in its own right, so a thorough understanding of its nature is basic to Artificial Intelligence.

This paper explores temporal explanation in the context of the situation calculus [McCarthy & Hayes, 1969], a formalism which, as well as being the oldest and best-understood logic-based formalism for representing change in AI, has considerable expressive power [Gelfond *et al*., 1991], [Lin & Shoham, 1992]. In this paper, two styles of representation for explanation problems are compared: the standard style used in the existing literature, and an alternative style. In addition, two fundamentally different approaches to explanation are explored: the deductive approach and the abductive approach. The paper presents the standard and alternative styles of representation first, then looks at the deductive approach, using both styles, and finally investigates the abductive approach.

Most attempts to formalise temporal explanation have adopted the deductive approach [Morgenstern & Stein, 1988], [Lifschitz & Rabinov, 1989], [Baker, 1989], [Crawford & Etherington, 1992]. Suppose we have a formula T which captures the timeless laws of change in a given domain, and a formula H representing when certain time-varying facts are true. According to the deductive approach, the explanation of an additional such fact F will be among the logical consequences of $T \wedge H \wedge F$. According to the abductive approach [Shanahan, 1989], an explanation is a formula Δ such that $T \wedge H \wedge \Delta$ has F among its logical consequences.

Throughout this paper, I will use the so-called stolen car problem (SCP) as a benchmark [Kautz, 1986]. The task is to model the reasoning involved in the following story. Suppose I park my car in the morning and go to work. At lunch time, I might reasonably apply default persistence and infer that the car is still where I left it. However, when I return to the car park in the evening I find that it has gone. Its disappearance requires an explanation. That is to say, we want to reason backwards in time to the (possible) causes of the car's disappearance. In this case, the only reasonable explanation for the car's disappearance is that it was stolen some time between morning and evening. So my previous conclusion that the car was still there at lunch time is open to question. The car may have been stolen any time after I parked it and before I observed that it was gone, so I cannot say anything about its whereabouts at lunch time.

1. Representing Explanation Problems

The ontology of the situation calculus includes situations, actions and fluents. A situation is an instantaneous snapshot of the world, and a fluent is anything whose value is subject to change. I will employ variables of three sorts corresponding to this ontology.[1] I will write Result(a,s) to denote the situation which results when action a is performed in situation s, and Holds(f,s) to represent that fluent f holds in situation s. If a fluent holds in a situation then it has the value true, and if it does not hold it has the value false. Several authors have attempted to deal with temporal explanation within the framework of the situation calculus [Lifschitz & Rabinov, 1989], [Baker, 1989], [Crawford & Etherington, 1992]. But I will now argue that the style of representation they all use is problematic.

To represent a particular domain using the situation calculus, we write two sets of sentences, one set describing which fluents change value as a result of performing each action (so-called axioms of motion), and one set describing which retain their value (so-called frame axioms). The main concern of a great deal of research on the formal representation of change has been the frame problem, or how to eliminate the need to write explicit frame axioms. One of the most successful attempts to overcome the frame problem is Baker's [1989].[2] His solution does not suffer from the difficulties pointed out by Hanks and McDermott [1987] and correctly handles ramifications (derived properties). It can also

[1] In what follows, variables begin with lower-case letters, and predicate and function symbols begin with upper-case letters. All variables are universally quantified unless otherwise indicated. A suitable set of uniqueness-of-names axioms is assumed.

[2] To follow closely the argument of this paper, the reader will require some familiarity with Baker's work.

cope with certain explanation problems. In particular, Baker represents the stolen car scenario as follows,

$$\neg\text{Holds(Stolen,S0)} \qquad \text{(SR1)}$$

$$\text{S2} = \text{Result(Wait,Result(Wait,S0))} \qquad \text{(SR2)}$$

$$\text{Holds(Stolen,S2)} \qquad \text{(SR3)}$$

Does this constitute a good representation of the SCP? Let us consider exactly what knowledge we are trying to capture. The meaning of Result(Wait,Result(Wait,S0)) is the situation which results when two successive Wait actions are performed in situation S0. The assertion that S2 equals this situation means that the only two actions which occur between S0 and S2 are the two Wait actions. It is implicit in this assertion that nothing else happens between S0 and S2. However, the whole point of the SCP is that we do not know what actions take place between S0 and S2. We don't know what S2 equals in terms of the Result function. Since the intended meaning of Wait is an action which has no effect, then it doesn't seem likely that S2 equals Result(Wait,Result(Wait,S0)).[3] However, since it is only by default that waiting has no effect, it is still possible to conclude that one of the wait actions is responsible for the car's disappearance.

Rather than half-heartedly asserting that nothing happens between S0 and S2 and allowing default reasoning to override this assertion to conclude that Wait actions sometimes have strange effects, a more intuitive representation of the SCP asserts nothing about S2 beyond the fact that it is the result of a sequence of actions which starts in situation S0. Then the aim of explanation is to characterise S2 in terms of the result function, that is to characterise the sequence of actions which starts in S0 and leads to a situation S2 in which the car is gone. Accordingly, I suggest the following representation of the SCP,[4]

$$\text{Holds(Car-parked,S0)} \qquad \text{(AR1)}$$

$$\neg\text{Holds(Car-parked,S2)} \qquad \text{(AR2)}$$

$$\text{Follows(S2,S0)} \qquad \text{(AR3)}$$

where Follows is defined thus,

$$\begin{aligned}&\text{Follows(sc,sa)} \leftrightarrow [\text{sc=sa} \vee \qquad \text{(AR4)}\\ &\quad \exists\text{a,sb [sc=Result(a,sb)} \wedge \text{Follows(sb,sa)]]}\end{aligned}$$

and where we have the following axiom of motion,

$$\neg\text{Holds(Car-parked,Result(Steal,s))} \qquad \text{(AR5)}$$

The point being made here applies to explanation using the situation calculus in general, and is not restricted to the SCP. Lifschitz and Rabinov [1989], for example, use the same style as Baker to represent a bloodless variation of the Yale shooting problem [Hanks & McDermott, 1987], in which the victim remains alive after the shooting. Their approach to explanation introduces the idea of a *miracle*, which is an unexpected effect of an action. Once again, in their approach default reasoning is expected to override a "half-hearted" assertion that nothing happens between loading and shooting to conclude that in fact the Wait action unloads the gun. As before, I suggest that the task of explanation is to determine exactly what sequence of actions takes place between loading and shooting.

In what follows, the style of representation exemplified by [Baker, 1989] and [Lifschitz & Rabinov, 1989] will be called the *standard* style, and the style which I have suggested will be called the *alternative* style. I will now examine both styles of representation in the context of the deductive approach to explanation, and later will examine both styles in the context of the abductive approach.

2. Deductive Approach, Standard Style

Underlying the deductive approach to explanation championed by Morgenstern and Stein [1988], Lifschitz and Rabinov [1989], Baker [1989], and Crawford and Etherington [1992] is a deductive approach to the assimilation of knowledge. Let us suppose that we have a formula T which represents an agent's knowledge about the world. Then, if the agent learns that F is the case, where F is not a consequence of T, the deductive approach to assimilating F is simply to add it to T. The formula $T \wedge F$ then represents the agent's knowledge about the world.

Using this approach, how is the SCP tackled within the framework of the situation calculus? Let's consider the standard style of representation first. As well as (SR1) to (SR3), we need a frame axiom. A common frame axiom is,

$$[\text{Holds(f,s)} \leftrightarrow \text{Holds(f,Result(a,s))}] \leftarrow \neg\text{Ab(a,f,s)} \qquad (1)$$

The frame problem is normally overcome by minimising the extension of Ab in some way, using circumscription for example. In Baker's work [1989], this is achieved by introducing an "existence-of-situations" axiom, then circumscribing, minimising Ab and allowing the Result function to vary. This avoids the problem Hanks and McDermott encountered with McCarthy's formulation [McCarthy, 1986], [Hanks & McDermott, 1987]. However, since the SCP doesn't involve actions with preconditions, it doesn't run into the Hanks-McDermott problem, and McCarthy's formulation, which minimises Ab and allows Holds to vary, is adequate.

Initially, we know just (SR1) and (SR2). With Wait the only action in the domain of discourse, nothing is abnormal, so minimising Ab using either McCarthy's or Baker's technique yields simply ¬Ab(a,f,s), from which we can conclude ¬Holds(Stolen,S2).

Using the deductive approach to explanation, when we learn (SR3) we simply add it to (SR1), (SR2) and (1), and derive a new set of conclusions. From (SR1) to (SR3) and (1), Baker [1989] gets,

$$\begin{aligned}&\text{Ab(Wait,Stolen,S0)} \vee\\ &\quad \text{Ab(Wait,Stolen,Result(Wait,S0))}\end{aligned}$$

This seems to be the consequence we intuitively expect, using the standard style of representation: the car is either stolen during the first Wait action or during the second, and we cannot say for sure which of these disjuncts is true. Minimising Ab simply reduces the set of models to those in which one of the disjuncts is true, the other one false, and Ab is false for everything else. However, this consequence

[3] In fact, the very idea of a "wait" action seems rather strange, and the idea of a sequence of two wait actions seems stranger still. Surely waiting is a pause between actions rather than an action in its own right.

[4] I prefer the formula Ý Holds(Car-parked,s) to Baker's Holds(Stolen,s), although they fulfil the same rôle.

doesn’t really constitute an explanation at all. It simply says that one of the Wait actions must have been abnormal. From (1), it can be seen that the abnormality of a Wait action is not sufficient to bring about a change in the value of Stolen. It is a necessary condition of such a change, not a sufficient one.

Furthermore, if the domain is widened a little, other difficulties arise. Suppose the domain includes actions with preconditions, thus necessitating a form of minimisation different to McCarthy’s. The best-known candidates at present are chronological minimisation (for example [Shoham, 1988]), causal minimisation (for example [Lifschitz, 1987]), and Baker’s state-based minimisation [1989]. As Baker points out, chronological minimisation, which postpones change until as late as possible, will insist that the car is stolen during the second Wait action; causal minimisation can be modified to cope with explanation [Lifschitz & Rabinov, 1989], but has problems with ramifications (derived properties); and his own approach, whilst adequate for the simple version of the problem presented above, falls apart as soon as another fluent is introduced which holds in S0.

Why should the need to tackle explanation problems interfere with our efforts to overcome the frame problem? In a later section, I will discuss the abductive approach to explanation, which doesn’t interfere with minimisation in any way, but first I will examine the deductive approach applied to the alternative style of representation suggested in Section 1.

3. Deductive Approach, Alternative Style

What happens when the deductive approach to explanation is used with the alternative style of representation? From (AR1) to (AR4) and (1), we have,

$$\exists a,sa,sb\ [Ab(a,Car\text{-}parked,sa) \wedge sb=Result(a,sa) \wedge Follows(sa,S0) \wedge Follows(S2,sb)]$$

From (AR5) and (1), minimising Ab using either McCarthy’s or Baker’s appraoch, we have,

$$Ab(a,f,s) \leftrightarrow [a=Steal \wedge f=Car\text{-}parked \wedge Holds(Car\text{-}parked,s)]$$

and therefore,

$$\exists sa,sb\ [sb=Result(Steal,sa) \wedge Follows(sa,S0) \wedge Follows(S2,sb)]$$

In other words, there is a Steal action between situations S0 and S2, which is the intuitively correct explanation. To simplify sentences of the above form, I introduce a new predicate. The formula Between(a,s1,s2) represents that an action a occurs between situations s1 and s2, and is defined as follows.

$$Between(a,sa,sd) \leftrightarrow \exists sb,sc\ [sc=Result(a,sb) \wedge Follows(sb,sa) \wedge Follows(sd,sc)] \qquad (AR6)$$

Then, the above explanation of the car’s disappearance can be abbreviated to,

$$Between(Steal,S0,S2)$$

So the deductive approach to the SCP seems to work using the alternative representation. Unlike the standard representation, the alternative representation doesn’t encounter difficulties with explanation problems in richer domains. Suppose that we employ Baker’s approach to minimisation — the Result function is allowed to vary, and there is an axiom asserting, for all possible combinations of fluents, the existence of a situation in which that combination holds. The problem that Baker reports [1989] using the standard representation is that the assertion that the car is not in the car park in S2 forces a new abnormality. There is a variety of choices for this abnormality, each of which satisfies Axiom (1) whilst allowing the car to disappear. Unfortunately, in a domain of any complexity, some of them are both counter-intuitive and minimal.

With the alternative representation, using Baker’s approach to minimisation, this problem simply doesn’t arise. The assertion that the car is not in the car park in S2 does not force a new abnormality. Rather, it forces a Steal action to occur between S0 and S2, and Steal actions are abnormal with respect to Car-parked anyway. So the minimisation of Ab is unaffected.

However, the approach described here is not complete without further minimisation. In the absence of (AR2), the explicit assertion that the car is not in the car park in S2, we wanted to be able to assume by default that it still was. From (AR1) and (AR3) to (AR6), knowing nothing about the theft, we wanted to be able to conclude Holds(Car-parked,S2). Unfortunately, (AR3) is too weak to allow this conclusion. It simply says that there is some sequence of actions between S0 and S2, and does not disallow the possibility of a Steal action occurring.

The alternative style of representation for explanation problems presupposes a framework which can cope with sequences of actions about which not everything is known. In the SCP, for example, we don’t know what actions have taken place between S0 and S2. However, we would like to assume by default that nothing happens we don’t know about.

There are several ways to achieve this. The approach I will sketch here is based on the work of Pinto and Reiter [1993]. The formula Actual(s) represents that the situation s is part of an actual narrative of events, about which we may have incomplete information. So we have, in the SCP example,

$$Actual(S0) \qquad Actual(S2)$$

The actual narrative of events corresponds to one path through the tree of situations defined by the Result function. The following three axioms guarantee this, following Pinto and Reiter [1993].

$$Actual(Result(a,s)) \rightarrow Actual(s)$$

$$[Actual(Result(a1,s)) \wedge Actual(Result(a2,s))] \rightarrow a1=a2$$

$$Result(a1,s1) = Result(a2,s2) \rightarrow [a1=a2 \wedge s1=s2]$$

A fourth axiom[5] is required to ensure that Baker's approach continues to work in the presence of the last of the above axioms.

$$[\forall f1[Holds(f1,s1) \leftrightarrow Holds(f1,s2)] \wedge Ab(a,f2,s1)] \rightarrow Ab(a,f2,s2)$$

[5] Thanks to Vladimir Lifschitz for suggesting this axiom.

Now Actual is minimised with a lower priority than Ab, and situation constants are allowed to vary, along with the Result function and the predicates Between and Follows. From now on, I will assume this new circumscription policy whenever I use the alternative style of representation. For further details the reader is referred to Pinto and Reiter [1993]. An alternative method for dealing with incomplete narratives, which could also be used here, is presented in [Miller & Shanahan, 1993].[6]

4. Preconditions

To complete the picture for the deductive approach with the alternative style, I will briefly investigate its application to an explanation problem involving preconditions. Consider (AR1) to (AR4) and (AR6), but suppose that it is a precondition of a successful theft that the car park is unguarded. So instead of (AR5) we have,

$$\neg \mathrm{Holds}(\mathrm{Car\text{-}parked},\mathrm{Result}(\mathrm{Steal},s)) \leftarrow \neg \mathrm{Holds}(\mathrm{Guarded},s) \qquad \text{(AR7)}$$

Initially the car park is guarded. But if a lazy security guard comes on duty, he immediately falls asleep, leaving the car park vulnerable to theft. We also know that Fred is a lazy security guard. To represent this, the action Guard(x) is introduced, denoting that security guard x comes on duty, along with the unary predicate Lazy.

$$\mathrm{Holds}(\mathrm{Guarded},S0) \qquad \text{(AR8)}$$

$$\mathrm{Holds}(\mathrm{Guarded},\mathrm{Result}(\mathrm{Guard}(x),s)) \leftrightarrow \neg \mathrm{Lazy}(x) \qquad \text{(AR9)}$$

$$\mathrm{Lazy}(\mathrm{Fred}) \qquad \text{(AR10)}$$

Now what can we conclude from the fact that the car is not parked in S2? The only plausible explanation, given the knowledge we have, is that Fred came on duty and fell asleep, leaving the car park unguarded. Then the car was stolen. Minimising Ab according to Baker's approach, we get,

$$\exists sa,sb\ [sb=\mathrm{Result}(\mathrm{Steal},sa) \wedge \mathrm{Follows}(sa,S0) \wedge \mathrm{Follows}(S2,sb) \wedge \neg \mathrm{Holds}(\mathrm{Guarded},sa)]$$

Then, working on the Holds conjunct of this formula, we can show,

$$\exists sa,sb,sc,sd,x\ [sb=\mathrm{Result}(\mathrm{Steal},sa) \wedge \mathrm{Follows}(sa,S0) \wedge \mathrm{Follows}(S2,sb) \wedge sd=\mathrm{Result}(\mathrm{Guard}(x),sc) \wedge \mathrm{Follows}(sc,S0) \wedge \mathrm{Follows}(sa,sd) \wedge \mathrm{Lazy}(x)]$$

which simplifies to,

$$\exists s,x\ [\mathrm{Between}(\mathrm{Guard}(x),S0,s) \wedge \mathrm{Between}(\mathrm{Steal},s,S2) \wedge \mathrm{Lazy}(x)]$$

In other words, a lazy security guard comes on duty and then the car is stolen. This is very nearly the desired result, but not quite because no mention is made of Fred, the only lazy security guard we know of. Of course, in a sense, this is quite correct, since nowhere have we said explicitly that Fred is the only lazy guard. On the other hand, if it was Fred that came on duty, that would explain the fact that the car park was unguarded at the wrong time.

To see that this could be a serious shortcoming, let's introduce a further complication to the story. In addition to the car park's being unguarded, there is another precondition to a successful theft. The alarm mustn't be on. Instead of (AR5) or (AR7), we have,

$$\neg \mathrm{Holds}(\mathrm{Car\text{-}parked},\mathrm{Result}(\mathrm{Steal},s)) \leftarrow \neg \mathrm{Holds}(\mathrm{Guarded},s) \wedge \neg \mathrm{Holds}(\mathrm{Alarm},s) \qquad \text{(AR11)}$$

Initially the alarm is indeed off, but if Fred comes on duty he always turns it on, knowing he's likely to fall asleep. However, if the thief smashes the alarm, it isn't on.

$$\neg \mathrm{Holds}(\mathrm{Alarm},S0) \qquad \text{(AR12)}$$

$$\mathrm{Holds}(\mathrm{Alarm},\mathrm{Result}(\mathrm{Guard}(\mathrm{Fred}),s)) \qquad \text{(AR13)}$$

$$\neg \mathrm{Holds}(\mathrm{Alarm},\mathrm{Result}(\mathrm{Smash},s)) \qquad \text{(AR14)}$$

The deductive approach cannot supply a more detailed explanation, in the light of these extra facts, than the one already given — a lazy security guard came on duty and then the car was stolen. Since it cannot be concluded that Fred was the lazy security guard who came on duty, using the deductive approach, we completely miss the subtlety that if it was Fred who came on duty, then the thief must have smashed the alarm.

Of course, it's true that "explanations come to an end somewhere," but this seems a little premature. We would like to find an approach to explanation that tells us that the following sequence of actions explains the car's disappearance — Fred came on duty, the alarm was smashed, and the car was stolen.

5. The Abductive Approach

Abduction is widely considered to be a mode of reasoning fundamental to AI, and it has a diverse range of applications, including diagnosis, planning, plan recognition, natural language interpretation, default reasoning, and of course temporal explanation. According to the abductive approach to explanation in the situation calculus, given a theory T comprising axioms of motion and the frame axiom (and any other necessary general axioms, such as Baker's "existence of situations"), and a history H representing that certain fluents hold in certain situations, to explain a new fact F representing that a fluent holds in a given situation we need to find a formula Δ such that $T \wedge H \wedge \Delta$ has F among its logical consequences.

In order to avoid trivial or weak explanations, a certain set of predicates are distinguished as *abducible*. Explanations have to be in terms of abducible predicates. Furthermore, to overcome the frame problem, some form of minimisation will be required. So more precisely, we say that, given T and H as above, a formula Δ is an explanation of a fact F if it mentions only abducible predicates, and $\mathrm{CIRC}[T \wedge H \wedge \Delta; P^*; Q^*] \models F$,[7] where P* and Q* are sets of predicates corresponding to a suitable circumscription policy to overcome the frame problem. Of course, there may be many such Δ's to explain any given fact. It is also convenient to

[6] The issue of narratives is orthogonal to the main point of the paper. The sketch given here is only offered as evidence that a working technique can be found.

[7] CIRC[y; P*; Q*] denotes the circumscription of the formula y minimising P* and allowing Q* to vary.

avoid explanations which are subsumed by other explanations. So we say that, given T and H, an explanation Δ of F is *minimal* if there is no explanation of F which is a subset of Δ.

In these abductive terms, what is the general form of an explanation problem expressed in the situation calculus? We are usually required to explain a conjunction of positive or negative Holds literals. Let's consider the SCP, using the standard style of representation first. We want to explain (SR3), and we require explanations in terms of previously unsuspected abnormalities. So the obvious policy is to make Ab abducible.

Let T be (1) and H be (SR1) $\wedge$ (SR2). Let Δ be Ab(Wait,Stolen,S0), and assume either McCarthy's or Baker's circumscription policy. As pointed out in Section 2, the abnormality of one of the Wait actions is a necessary but not a sufficient condition for the car to be stolen. Appropriately then, Δ is not an explanation of (SR3) at all according to the abductive approach. Similarly, if we let Δ be Ab(Wait,Stolen,Result(Wait,S0)), then it is still no explanation. In fact, given the standard representation and the abductive approach with Ab made abducible, the disappearance of the car literally defies explanation. Furthermore, since it incorporates no knowledge of Steal actions, the standard representation doesn't permit any explanation of the car's disappearance without the inclusion in Δ of new axioms of motion.

Now let's consider the alternative style. The explanations we require are in terms of the sequence of actions which takes place between two situations. So the obvious abduction policy is to make Between abducible. In the SCP, we want to explain (AR2). Let T be the conjunction of (1) and (AR4) to (AR6), and let H be (AR1) $\wedge$ (AR3). Suppose we minimise abnormality according to either McCarthy's or Baker's approach, and we also minimise Actual. Consider Δ=Between(Steal,S0,S2). Does this constitute an explanation?

Minimising Actual yields S2=Result(Steal,S0). Then, applying (AR5), we have $\neg$Holds(Car-parked,S2). So Δ is indeed an explanation. There are other explanations too, but each of these involves a sequence of Steal actions. It is easy to see that Δ subsumes all of these explanations, and therefore all minimal explanations will be equivalent to Δ. This approach bears a strong similarity to that of Green [1969] and Kowalski [1979, Chapter 6] to plan formation in the situation calculus, in which resolution generates a binding of the form s=Result(a1,Result(a2,...)) to solve a goal of the form Holds(f,s). This binding conforms exactly to the abductive idea of an explanation with the alternative style of representation, where equality is made abducible.

Note that if we asserted that another action, say going to lunch, occurred between S0 and S2, then this Δ would still constitute an explanation, and would furthermore be neutral about the relative order of lunch time and the car's theft. So it would not be possible to conclude, in the presence of Δ, that the car was still in the car park at lunch time.

Next, we'll examine how the abductive approach fares with the alternative style of representation with preconditions. Recall the variant of the SCP with Fred, the lazy security guard who switches on the alarm when he comes on duty. Once again, we want to explain (AR2). This time, assume Baker's minimisation technique, to ensure that the precondition is properly treated. Let T be the conjunction of an existence of situations axiom with (1), (AR4), (AR6), (AR9) to (AR11), (AR13) and (AR14). Let H be the conjunction of (AR1), (AR3), (AR8) and (AR12). Let Δ be,

$$\exists \text{sa,sb [Between(Guard(Fred),S0,sa)} \wedge$$
$$\text{Between(Smash,sa,sb)} \wedge \text{Between(Steal,sb,S2)]}$$

Let S = Result(Smash,Result(Guard(Fred),S0)). The minimisation of Actual now gives S2 = Result(Steal,S). Applying (AR9) and (AR10), we get $\neg$Holds(Guarded,S). We get Holds(Alarm,Result(Guard(Fred),S0)) by applying (AR13), but by applying (AR14) we get $\neg$Holds(Alarm,S). Finally, applying (AR11) we get $\neg$Holds(Car-parked,S2). So Δ is an explanation. Again there are other explanations, involving sequences of Steal, Guard and Smash actions, and again these are all subsumed by Δ, so any minimal explanation will be equivalent to Δ.

By way of contrast, the closest thing to an explanation supplied by the deductive approach, namely

$$\exists \text{s,x [Between(Guard(x),S0,s)} \wedge$$
$$\text{Between(Steal,s,S2)} \wedge \text{Lazy(x)]}$$

doesn't constitute an explanation at all according to the abductive approach, even if we make Lazy abducible. This is because it ignores the possibility that the lazy security guard is Fred, who will turn the alarm on, thus preventing the Steal action from being successful.

Discussion

This paper is intended to be a critical study of various approaches to explanation within the framework of the situation calculus. The analysis would seem to recommend the abductive approach with the alternative style of representation. However, a number of issues remain to be discussed.

To begin with, the paper has adopted the situation calculus, with circumscription as a means of default reasoning, and has employed Baker's approach to the frame problem. There are, of course, many alternatives. However, I conjecture that the lessons learned here will apply to other formalisms, other forms of default reasoning, and other approaches to the frame problem (see [Shanahan, 1989], for example).

The impression given in this paper is that abduction and deduction are competing approaches to explanation. But it could be argued that abduction isn't a particular approach to explanation, it is the nature of explanation. A particular approach to explanation might perform abduction directly, or it might simulate it through deduction, so long as the explanations it produced conformed to the abductive definition. Under this interpretation, there is no need to show the adequacy of the abductive approach, because it supplies the very criterion of adequacy.

A related issue which merits some discussion is that of knowledge assimilation. A problem like the stolen car problem can be thought of simply as a reasoning problem — what are the possible explanations of the car's disappearance. Alternatively, it can be thought of as a knowledge assimilation problem — how is the fact of the car's disappearance to be assimilated. The abductive and deductive approaches to explanation imply different views of knowledge assimilation. Suppose that we have a knowledge base in the form of a formula T. Under a classical, deductive view of knowledge assimilation, new facts are always added directly to T. With an abductive view of knowledge assimilation, not every fact is eligible for direct addition to T. Sometimes the assimilation of a new fact F demands the addition of a formula Δ of a certain form to T such that $T \wedge \Delta \models F$ [Kowalski, 1979, Chapter 13]. That is, new facts sometimes have to be explained through abduction.

Using abduction with the situation calculus, assimilating a new Holds fact, such as the fact that my car is not in the car park in the evening, demands the addition of a formula representing that certain actions take place, so that the new fact becomes a logical consequence of the knowledge base. With the stolen car problem, there is a unique minimal explanation, but this not necessarily the case. One approach to dealing with multiple explanations is to add the disjunction of all minimal explanations to the knowledge base, but this issue is beyond the scope of this paper.

Two important questions come to mind here. Why do some facts demand explanation when others do not? And why are some predicates abducible when others are not? In so far as a problem like the SCP is viewed simply as a reasoning problem, these questions are not very important, since the answers have to be written into the specification of the problem. But taking the wider, knowledge assimilation view, the questions become more pressing. A simple answer is that anything which can be considered a first cause doesn't require explanation, whereas anything which cannot be considered a first cause does require explanation. For example, we might decide to consider the occurrence of an action as a first cause, but not the effects of an action. This is a partial justification for making Between abducible, and insisting that Holds facts, except those about the initial situation, must be explained. Clearly though, these issues merit further study.

Finally, an important question is the relationship between abduction and deduction [Console *et al.*, 1991], [Konolige, 1992]. When do they coincide? Or, if abduction is adopted as the specification of explanation as suggested above, when does deduction conform to that specification? And why does abduction work in some cases when deduction doesn't? In essence, abduction finds sufficient conditions for a fact to hold, whilst deduction only finds necessary conditions. In certain circumstances, necessary conditions are also sufficient conditions. This is the case when the knowledge involved is expressed in terms of biconditionals. The frame axiom (1), for example, makes it a necessary and sufficient condition for a fluent to hold in Result(a,s) that the fluent holds in s, given that a isn't abnormal in this context. Furthermore, one-way implications can sometimes behave like biconditionals in this way when minimisation is involved, because minimisation often has the effect of "completing" the implication, that is turning it into a biconditional. This was the case with Ab in the SCP. However, there is no reason to suppose that necessary and sufficient conditions will always coincide, even in the presence of minimisation, as we saw with extended SCP, in which deduction failed because the predicate Lazy was not completed.

Acknowledgements

Thanks to Tony Kakas, Bob Kowalski, Vladimir Lifschitz, Rob Miller and Javier Pinto. The author is supported by an SERC Advanced Research Fellowship.

References

[Baker, 1989], A.B.Baker, A Simple Solution to the Yale Shooting Problem, *Proceedings KR 89*, p 11.

[Console *et al.*, 1991] L.Console, D.Dupré and P.Torasso, On the Relationship between Abduction and Deduction, *Journal of Logic and Computation*, vol 1 (1991), p 661.

[Crawford & Etherington, 1992] J.M.Crawford and D.W.Etherington, Formalizing Reasoning about Change: A Qualitative Reasoning Approach, *Proceedings AAAI 92*, p 577.

[Gelfond *et al.*, 1991] M.Gelfond, V.Lifschitz and A.Rabinov, What Are the Limitations of the Situation Calculus? in *Essays for Bledsoe*, ed R.Boyer, Kluwer Academic (1991), p 167.

[Green, 1969] C.Green, Applications of Theorem Proving to Problem Solving, *Proceedings IJCAI 69*, p 219.

[Hanks & McDermott, 1987] S.Hanks and D.McDermott, Nonmonotonic Logic and Temporal Projection, *Artificial Intelligence*, vol 33 (1987), p 379.

[Kautz, 1986] H.Kautz, The Logic of Persistence, *Proceedings AAAI 86*, p 401.

[Konolige, 1992] K.Konolige, Abduction Versus Closure in Causal Theories, *Artificial Intelligence*, vol 53 (1992), p 255.

[Kowalski, 1979] R.A.Kowalski, Logic for Problem Solving, North Holland, 1979.

[Lifschitz, 1987] V.Lifschitz, Formal Theories of Action, *Proceedings of the 1987 Workshop on the Frame Problem*, p 35.

[Lifschitz & Rabinov, 1989] V.Lifschitz and A.Rabinov, Miracles in Formal Theories of Action, *Artificial Intelligence*, vol 38 (1989), p 225.

[Lin & Shoham, 1992] F.Lin and Y.Shoham, Concurrent Actions in the Situation Calculus, *Proceedings AAAI 92*, p 590.

[McCarthy, 1986] J.McCarthy, Applications of Circumscription to Formalizing Common Sense Knowledge, *Artificial Intelligence*, vol 26 (1986), p 89.

[McCarthy & Hayes, 1969] J.McCarthy and P.J.Hayes, Some Philosophical Problems from the Standpoint of Artificial Intelligence, in *Machine Intelligence 4*, ed D.Michie and B.Meltzer, Edinburgh University Press (1969).

[Miller & Shanahan, 1993] R.S.Miller and M.P.Shanahan, Narratives in the Situation Calculus, Imperial College Department of Computing Technical Report.

[Morgenstern & Stein, 1988] L.Morgenstern and L.A.Stein, Why Things Go Wrong: A Formal Theory of Causal Reasoning, *Proceedings AAAI 88*, p 518.

[Pinto & Reiter, 1993] J.Pinto and R.Reiter, Temporal Reasoning in Logic Programming: A Case for the Situation Calculus, to appear in *Proceedings ICLP 93*.

[Shanahan, 1989] M.P.Shanahan, Prediction Is Deduction but Explanation Is Abduction, *Proceedings IJCAI 89*, p 1055.

[Shoham, 1988] Y.Shoham, Reasoning About Change: Time and Change from the Standpoint of Artificial Intelligence, MIT Press (1988).