Why Things Go Wrong: A Formal Theory of Causal Reasoning

Leora Morgenstern and **Lynn Andrea Stein**
Department of Computer Science
Brown University
Box 1910, Providence, RI 02912

Abstract

This paper presents a theory of generalized temporal reasoning. We focus on the related problems of

1. Temporal Projection—determining all the facts true in a chronicle, given a partial description of that chronicle, and
2. Explanation—figuring out what went wrong if an unexpected outcome occurs.

We present a non-monotonic temporal logic based on the notion that actions only happen if they are *motivated.* We demonstrate that this theory handles generalized temporal projection correctly, and in particular, solves the Yale Shooting Problem and a related class of problems. We then show how our model lends itself to a very natural characterization of the concept of an adequate explanation for an unexpected outcome.

1 Introduction

A theory of generalized temporal reasoning is a crucial part of any theory of commonsense reasoning. Agents who are capable of tasks ranging from planning to story understanding must be able to predict from their knowledge of the past what will happen in the future, to decide on what must have happened in the past, and to furnish a satisfactory explanation when a projection fails.

This paper present a theory that is capable of such reasoning. We focus on the related problems of

1. Temporal Projection—determining all of the facts that are true in some chronicle, given a partial description of that chronicle, and
2. Explanation—determining what went wrong if an unexpected outcome occurs.

Most AI researchers in the area of temporal reasoning have concentrated their efforts on parts of the temporal projection task: in particular, on the problem of forward temporal projection, or prediction ([McCarthy and Hayes, 1969], [McDermott, 1982], [Hayes, 1985], [Shoham, 1987]). Standard logics are unsuitable for the prediction task because of such difficulties as the frame problem. Straightforward applications of non-monotonic logic to temporal logics (suggested by [McDermott, 1982], [McCarthy, 1980]) are also inadequate, as [Hanks and McDermott, 1986] demonstrated through the Yale Shooting Problem.

Several solutions to the Yale Shooting Problem, using extensions of default logic, have been proposed ([Shoham, 1986] [Shoham, 1987], [Kautz, 1986], [Lifschitz, 1986] [Lifschitz, 1987], [Haugh, 1987]). All of these solutions, however, while adequate for the Yale Shooting Problem itself, handle either forward or backward projection incorrectly, and/or work only within a very limited temporal ontology. Thus, they cannot serve as the basis for a theory of generalized temporal reasoning.

In this paper, we present a solution to the problems of both forward and backward temporal projection, based upon the concept that actions happen only if they have to happen. We then show how our model lends itself to a very natural characterization of the concept of an adequate explanation for an unexpected outcome.

In the next section, we survey the solutions that have been proposed to the YSP, and explain why they cannot handle general temporal projection accurately. We then present our theory of default temporal reasoning and demonstrate that it can handle the Yale Shooting Problem as well as the problems that give other theories difficulty. Finally, we extend our theory of temporal projection to a theory of explanation.

2 Previous Approaches to the Prediction Problem

2.1 Default Reasoning and the Yale Shooting Problem

The frame problem—the problem of determining which facts about the world stay the same when actions are performed—is an immediate consequence of the attempt to subsume temporal reasoning within first order logic. McCarthy and Hayes first discovered this problem when they developed the situation calculus ([McCarthy and Hayes, 1969]); however, it is not restricted to the situation calculus and in fact arises in all reasonably expressive temporal ontologies ([McDermott, 1987]). In order to deal with the frame problem, McCarthy and Hayes suggested using frame axioms to specify the facts that don't change when certain actions are performed; critics (*e.g.* [McDermott, 1984]) have argued that such an approach is unsatisfactory given the difficulty of writing such axioms, the intractability of a theory containing so many axioms, and the fact that frame axioms are often false. This last point is especially relevant for temporal ontologies which allow for concurrent actions.

[McDermott, 1982] introduced the notion of a persistence: the time period during which a property typically persists. He argued that we reason about what is true in the world, not via frame axioms, but through our knowledge of the persistences of various properties. Such reasoning is inherently non-monotonic.

These considerations led [McDermott, 1982] to argue that temporal reasoning is best formalized within a non-monotonic logic. The discovery of the Yale Shooting Problem ([Hanks and McDermott, 1986]), however, demonstrated that this might not always yield desirable results.

The Yale Shooting Problem can briefly be described as follows: Assume that a gun is loaded at time 1, and the gun is fired (at Fred) at time 5. We know that if one loads a gun at time j, it is loaded at time j+1[1] that if a loaded gun is fired at a person at time j, the person is dead at time j+1, that if a gun is loaded at j, it will typically be loaded at time j+1 ("loaded" persists for as long as possible), and that if a person is alive at time j, he will typically be alive at time j+1 ("alive" persists for as long as possible).

We would like to predict that Fred is dead at time 6. Relative to standard non-monotonic logics ([McDermott and Doyle, 1980], [McCarthy, 1980], [Reiter, 1980]), however, the chronicle description supports (at least) two models: the expected one, in which one reasons by default that the gun is loaded at time 5, and in which Fred is dead at time 6, and the unexpected model in which one reasons by default that Fred is alive at time 6, and in which, therefore, the gun must be unloaded at time 5. Standard non-monotonic logic gives us no way of preferring the expected, intuitively correct model to the unexpected model.

Like the frame problem, the Yale Shooting Problem was first presented within the situation calculus framework, but is not restricted to that particular ontology ([McDermott, 1987]).

2.2 Proposed Solutions to the YSP and Their Limitations

In their original discussion of the Yale Shooting Problem, Hanks and McDermott argued that the second, unexpected model seems incorrect because we tend to reason forward in time and not backward. The second model seems to reflect what happens when we reason backward. Such reasoning, they argued, is unnatural: the problem with non-monotonic logic is that there is no way of preferring the forward reasoning models to the backward reasoning models.

2.2.1 Chronological Minimization

The first wave of solutions to the Yale Shooting Problem ([Shoham, 1986], [Kautz, 1986], [Lifschitz, 1986]) all independently set out to prove that such a preference could indeed be expressed in non-monotonic logic. We discuss Shoham's work here: criticisms of his theory apply equally to the others in the group.

Shoham defines the following preference relation on models: $\mathcal{M}_1$ is preferable to $\mathcal{M}_2$ if $\mathcal{M}_1$ and $\mathcal{M}_2$ agree up to some time point j, but at j, there is some fact known to be true in $\mathcal{M}_2$, which is not known to be true in $\mathcal{M}_1$. $\mathcal{M}_1$ is said to be chronologically more ignorant than $\mathcal{M}_2$. This preference defines a partial order; models which are minimal elements under this ordering are said to be chronologically maximally ignorant.

The expected model—in which Fred is dead—is preferable to the unexpected model—in which Fred is alive, since, in the unexpected model, it would be known that at some point before 5, something happened to unload the gun. In fact, in all chronologically maximally ignorant models for this set of axioms, the gun is loaded at time 5, and therefore, Fred is dead.

Solutions based upon forward reasoning strategies have two drawbacks. In the first place, agents perform both backward and forward reasoning. In fact, agents typically do backward reasoning when performing backward temporal projection. Consider, for example, a modification of the Yale Shooting Problem, where we are told that Fred is alive at time 6. We should know that the gun must somehow have become unloaded between times 2 and 5; however, we should not be able to say exactly when this happened. In contrast to this intuition, the systems of Shoham and Kautz would predict that the gun became unloaded between time 4 and time 5. This is because things stay the same for as long as possible.[2]

A second objection to the strategy of chronological minimization is that it does not seem to address the real concerns underlying the Yale Shooting Problem. We don't reason that Fred is dead at time 6 *because* we reason forward in time. We conclude that Fred is dead because we are told of an action that causes Fred's death, but are not told of any action that causes the gun to be unloaded.

2.2.2 Circumscribing Over Causes

[Lifschitz, 1987] and [Haugh, 1987] independently proposed solutions which were not based upon forward reasoning strategies. We present Lifschitz's; again criticisms of his theory apply to both. Lifschitz's solution is based on the intuition that "all changes in the values of fluents are caused by actions." Lifschitz introduces a predicate causes(act,f,v), where action act causes fluent f to take on value v, and a predicate precond(f,act). Success is defined in terms of precond, affects in terms of causes and success. He circumscribes over both the causes and precond predicates; circumscribing over causes solves the frame problem.[3] Things are only caused when there are axioms implying that they are caused. Necessary preconditions for an action are satisfied only when the axioms force this to be the case. Actions are successful exactly when all preconditions hold; actions affect the values of fluents if and only if some successful action causes the value to change. Assuming, now, the following axioms: causes(load,loaded,true), causes(shoot,loaded,false), causes(shoot,alive,false), precond(loaded,shoot), and a chronicle description stating that a load takes place at 1, a wait at 2,3, and 4, and a shoot at 5, we can predict that Fred is dead at time 6. There is no way that the wait action can cause the fluent loaded to take on the value false.

This solution doesn't force reasoning to go forward in time. Nevertheless, Lifschitz's solution is highly problematic. It works only within rigid formalisms like the situation calculus, and cannot be extended to—and in fact yields incorrect results in—a more flexible, realistic theory.

[1] It is implicitly assumed that actions take unit time.

[2] This point was noted by Kautz when he first presented his solution to the Yale Shooting Problem.

[3] Lifschitz introduces the precond predicate in order to solve the qualification problem, which we don't discuss here.

Moreover, a closer examination of the solution shows that it does not address one of the major intuitions underlying the Yale Shooting Problem.

It is crucial to realize that the causes predicate over which Lifschitz circumscribes ranges over action types as opposed to action instances. Circumscribing over causes thus entails that state changes will not happen spontaneously, but does not in general entail that as little will change as possible. Since the situation calculus framework itself entails that as little as possible happens, the solution will work as long as we stay within this rigid framework. Problems arise, however, in frameworks in which not all actions are known.

Consider what would happen in a world in which concurrent actions were allowed, and in which we were to add the rule causes(unload,loaded,false) to the theory. We could then have a model $\mathcal{M}_1$ where an unload occurs at time 2, the gun is thus unloaded, and Fred is alive at time 6. There would be no way to prefer the expected model where Fred dies to this model.[4] This cannot in fact happen in Lifschitz's formulation because in the situation calculus, concurrent actions aren't allowed. Since a wait action occurs at times 2, 3, and 4, nothing else can occur, and unload actions are ruled out.

Lifschitz's solution thus works only in frameworks where all the acts in a chronicle are known. In these cases, circumscribing the cause predicate gives us exactly what we want—it disables spontaneous state changes. The intuition underlying the Yale Shooting Problem, however, is that we can make reasonable temporal projections in worlds where concurrent actions are allowed, even if we aren't necessarily told of all the events that take place in a chronicle. The fact is that even if we are given a *partial* description, we will generally not posit additional actions unless there is a good reason to do so.

The temporal projection problem is thus a dual one: we must reason that actions don't cause fluents to take on values in unexpected ways, and we must reason that unexpected events don't in general happen. Lifschitz solved the first of these problems; in the next section, we turn our attention to the second.

3 Temporal Projection: A Theory of Motivated Actions

In this section, we develop a model of temporal projection which yields a satisfying solution to the Yale Shooting Problem, and which lends itself nicely to a theory of explanation. Our model formalizes the intuition that we typically reason that events in a chronicle happen only when they "have to happen". We formalize the idea of a *motivated action*, an action that *must* occur in a particular model.

3.1 The Formal Theory

We begin by formally describing the concepts of a theory and a chronicle description. We work in a first order logic L, augmented by a simple temporal logic. Sentences are of the form True(j,f) where t is a time point, and f is a fluent—a term representing some property that changes with time. True(j,¬f) iff ¬True(j,f). Occurs(act) and loaded are examples of fluents. If φ = True(j,f), j is referred to as the *time point* of φ, time(φ). Time is isomorphic to the integers. Actions are assumed to take unit time.

A *theory*, *T*, and a *chronicle description*, *CD*, are sets of sentences of L. The union of a theory and a chronicle description is known as a *theory instantiation*, *TI*. Intuitively, a theory contains the general rules governing the behavior of (some aspects of) the world; a chronicle description describes some of the facts that are true in a particular chronicle. A theory includes *causal rules* and *persistence rules*. A causal rule is a sentence of the form $\alpha \wedge \beta \Longrightarrow \gamma$, where:

α is a non-empty set of sentences of the form True(j,Occurs(act))—the set of *triggering events* of the causal rule,

β is a conjunction of statements stating the preconditions of the action, and

γ describes the results of the action.

Note that γ can include sentences of the form True(j+1,Occurs(act)). We can thus express causal chains of action.

A persistence rule is of the form

$$\text{True(j,p)} \wedge \beta \Longrightarrow \text{True(j+1,p)}$$

where β includes a conjunction of statements of the form

$$\text{True(j,}\neg\text{Occurs(act))}$$

These persistence rules bear a strong resemblance to frame axioms. In reality, however, they are simply instances of the principle of inertia: things do not change unless they have to.

We have hand coded the persistence rules for this simple case, although it is not necessary to do so. They can in fact be automatically generated from the theory's causal rules, relative to a closed world assumption on causal rules: that all the causal rules that are true are in the theory. This is indeed exactly what Lifschitz achieves by circumscribing over the causes predicate in his formulation. It is likely that such a strategy will be an integral part of any fully developed theory of temporal projection. Since the automatic generation of persistence rules is not the main thrust of this paper, we will not develop this here.

It is important to note that all of the rules in any theory *T* are monotonic. We achieve non-monotonicity solely by introducing a preference criterion on models: in particular, preferring models in which the fewest possible extraneous actions occur. Typically, we will not be given

[4] Haugh seems to address a related point in his paper. Haugh considers the case where we have an axiom stating that unload causes loaded to be false, and that the precondition for unload is that the performing agent knows how to perform the action (we recast into Lifschitz's formalism here for ease of comparison). Then, if it is known that an unload(attempt) Occurs, there will be no way of preferring models where loaded is true to models where loaded is false. Haugh says that this is to be expected; if we know of an unload attempt, we do not want to conclude that loaded is true. This argument is really beside the point. It is quite clear that if we are told of an unload (attempt), we will not conclude that Fred is dead. The point of the YSP is that, if you are not explicitly told of an unload, you will not seriously consider the possibility when making a prediction.

enough information in a particular chronicle description to determine whether or not the rules in the theory fire. However, because persistence rules explicitly refer to the non-occurrence of events, and because we prefer models in which events don't occur unless they have to, we will in general prefer models in which the persistence rules do fire. The facts triggered by persistence rules will often allow causal rules to fire as well.

3.2 Motivated Actions

Given a particular theory instantiation, we would like to be able to reason about the facts which ought to follow from the chronicle description under the theory. In particular, we would like to be able to determine whether a statement of the form True(j,p) is true in the chronicle. If j is later than the latest time point mentioned in *CD*, we call this reasoning *prediction*, or *forward projection*, otherwise, the reasoning is known as *backward projection*.

Given $TI = T \cup CD$, we are thus interested in determining the preferred models for *TI*. $\mathcal{M}(TI)$ denotes a model for *TI*: *i.e.*, $(\forall \varphi \in TI)[\mathcal{M}(TI) \models \varphi]$. We define a preference criterion for models in terms of *motivated* actions: those actions which "*have to happen.*" Our strategy will be to minimize those actions which are *not* motivated.

Definition: Given a theory instantiation $TI = T \cup CD$, we say that a statement φ is *motivated in* $\mathcal{M}(TI)$ if it is strongly motivated in $\mathcal{M}(TI)$ or weakly motivated in $\mathcal{M}(TI)$.

A statement φ is *strongly* motivated with respect to *TI* if φ is in all models of *TI*, *i.e.* if $(\forall \mathcal{M}(TI))[\mathcal{M}(TI) \models \varphi]$. If φ is strongly motivated with respect to *TI*, we say that it is motivated in $\mathcal{M}(TI)$, for all $\mathcal{M}(TI)$.

A statement φ is *weakly* motivated in $\mathcal{M}(TI)$ if there exists in *TI* a causal or persistence rule of the form $\alpha \wedge \beta \Longrightarrow \varphi$, α is (strongly or weakly) motivated in $\mathcal{M}(TI)$, and $\mathcal{M}(TI) \models \beta$.

Intuitively, φ is motivated in a model if it *has to be* in that model. Strong motivation gives us the facts we have in *CD* to begin with as well as their closure under *T*. Weak motivation gives us the facts that have to be in a *particular* model relative to *T*. Weakly motivated facts give us the non-monotonic part of our model—the conclusions that may later have to be retracted.

We now say that a model is preferred if it has as few unmotivated actions as possible. Formally, we define the preference relation on models as follows:

Definition: Let φ be of the form True(j,Occurs(act)). $\mathcal{M}_i(TI) \trianglelefteq \mathcal{M}_j(TI)$ ($\mathcal{M}_i$ is *preferable* to $\mathcal{M}_j$) if $(\forall \varphi)[\mathcal{M}_i(TI) \models \varphi \wedge \mathcal{M}_j(TI) \not\models \varphi \Longrightarrow \varphi$ is motivated in $\mathcal{M}_i(TI)$.

That is, $\mathcal{M}_i(TI)$ is preferable to $\mathcal{M}_j(TI)$ if any action which occurs in $\mathcal{M}_i(TI)$ but does not occur in $\mathcal{M}_j(TI)$ is motivated in $\mathcal{M}_i(TI)$. Note that such actions can only be weakly motivated; if an action is strongly motivated, it is true in all models.

Definition:
If both $\mathcal{M}_i(TI) \trianglelefteq \mathcal{M}_j(TI)$ and $\mathcal{M}_j(TI) \trianglelefteq \mathcal{M}_i(TI)$, we say that $\mathcal{M}_i(TI)$ and $\mathcal{M}_j(TI)$ are *equipreferable* ($\mathcal{M}_i(TI) \bowtie \mathcal{M}_j(TI)$).

$\trianglelefteq$ induces a partial ordering on acceptable models of *TI*. A model is *preferred* if it is a minimal element under $\trianglelefteq$:

Definition: $\mathcal{M}(TI)$ is a *preferred model* for *TI* if $\mathcal{M}'(TI) \trianglelefteq \mathcal{M}(TI) \Longrightarrow \mathcal{M}'(TI) \bowtie \mathcal{M}(TI)$.

Since not all models are comparable under $\trianglelefteq$, there may be many preferred models. Let $\mathcal{M}^*$(TI) be the union of all preferred models.

We define the following sets:

$\cap_{\mathcal{M}^*} = \{\varphi \mid (\forall \mathcal{M} \in \mathcal{M}^*(TI))[\mathcal{M} \models \varphi]\}$—the set of statements true in all preferred models of *TI*

$\cup_{\mathcal{M}^*} = \{\varphi \mid (\exists \mathcal{M} \in \mathcal{M}^*(TI))[\mathcal{M} \models \varphi]\}$—the set of statements true in at least one preferred model of *TI*

Consider, now, the relationship between any statement φ and *TI*. There are three cases:

Case I: φ is in $\cap_{\mathcal{M}^*(TI)}$. In this case, we say that *TI projects* φ.

Case II: φ is in $\cup_{\mathcal{M}^*(TI)}$. In this case, we say that φ is *consistent with TI*. However, *TI* does not project φ.

Case III: φ not in $\cup_{\mathcal{M}^*(TI)}$. In this case, we say that φ is *inconsistent with TI*. In fact, it is the case that *TI* projects $\neg\varphi$.

If *TI* projects φ, and time(φ) is later than the latest time point mentioned in *TI*, we say that *TI predicts* φ.

3.3 Prediction: The Yale Shooting Problem, Revisited

We now show that our theory can handle the Yale Shooting Problem. We represent the scenario with the following theory instantiation:
CD:

True(1,alive)
True(1,load)
True(5,shoot)

T contains the causal rules for shoot, load, and unload, as well as the persistences for loaded and alive:

T: Causal Rules:

$$\begin{array}{rcl}
\mathsf{True(j,Occurs(load))} & \Longrightarrow & \mathsf{True(j{+}1,loaded)} \\
\mathsf{True(j,Occurs(shoot))} & \wedge & \mathsf{True(j,loaded)} \\
 & \Longrightarrow & \mathsf{True(j{+}1,\neg alive)} \\
\mathsf{True(j,shoot)} & \Longrightarrow & \mathsf{True(j{+}1,\neg loaded)} \\
\mathsf{True(j,unload)} & \Longrightarrow & \mathsf{True(j{+}1,\neg loaded)}
\end{array}$$

Persistence Rules:

$$\begin{array}{rcl}
\mathsf{True(j,alive)} & \wedge & \mathsf{(True(j,\neg Occurs(shoot))} \\
 & & \vee\ \mathsf{True(j,\neg loaded))} \\
 & \Longrightarrow & \mathsf{True(j{+}1,alive)} \\
\mathsf{True(j,loaded)} & \wedge & \mathsf{True(j,\neg Occurs(shoot))} \\
 & \wedge & \mathsf{True(j,\neg Occurs(unload))} \\
 & \Longrightarrow & \mathsf{True(j{+}1,loaded)}
\end{array}$$

Let $\mathcal{M}_1$ be the expected model, where the gun is loaded at 5, and Fred is dead at 6; and let $\mathcal{M}_2$ be the unexpected model, where an unload takes place at some time between 2 and 5, and therefore Fred is alive at 6. Both $\mathcal{M}_1$ and $\mathcal{M}_2$ are models for *TI*. However, we will see that $\mathcal{M}_1$ is preferable to $\mathcal{M}_2$, since extra, unmotivated actions take place in $\mathcal{M}_2$.

We note that the facts True(1,alive), True(1,Occurs(load)), and True(5,Occurs(shoot)) are strongly motivated, since they are in *CD*. The fact True(2,loaded) is also strongly motivated; it is not in *CD*, but it must be true in all models of *TI*. In $\mathcal{M}_1$, the model in which the gun is still loaded at 5, True(6,¬alive) is weakly motivated. It is triggered by the shoot action, which is motivated, and the fact that the gun is loaded, which is true in $\mathcal{M}_1$. In $\mathcal{M}_2$, the occurrence of the unload action is unmotivated. It is not triggered by anything.

According to this definition, then, $\mathcal{M}_1$ is preferable to $\mathcal{M}_2$. There is no action which occurs in $\mathcal{M}_1$ that does not occur in $\mathcal{M}_2$. However, $\mathcal{M}_2$ is not preferable to $\mathcal{M}_1$: there is an action, unload, which occurs in $\mathcal{M}_2$, but not in $\mathcal{M}_1$, and this action is unmotivated.

In fact, it can be seen that in any preferred model of *TI*, the gun must be loaded at time 5, and therefore Fred must be dead at time 6. That is because in a model where the gun is unloaded at 5, a shoot or unload action must happen between times 2 and 5, and such an action would be unmotivated. Since the facts that loaded is true at time 5 and that Fred is dead at time 6 are in all preferred models of *TI*, *TI* projects these facts.

Note that preferring models in which the fewest possible unmotivated actions occur is not equivalent to preferring models in which the fewest possible actions occur. Consider, *e.g.*, a theory of message passing in which messages go through several checkpoints before completion. The message is passed as long as the control switch is open. An action is needed to close the switch. If we know that the message is started, we would like to predict that the switch remains open and the message completes. This is in fact what our preference criterion projects. However, since each stage of the message passing can be regarded as a separate action, a theory minimizing occurrences will predict that the switch is turned off, eliminating additional message passing segments.

3.4 Backward Projection

We now show that our theory handles backward projection properly. As an example, consider *TI*′, where $TI' = TI \cup$ {True(6,alive)}. That is, *TI*′ is the theory instantiation resulting from adding the fact that Fred is alive at time 6 to the chronicle description of *TI*. Since we know that a shoot occurred at 5, we know that the gun cannot have been loaded at 5. However, we also know that the gun was loaded at 2. Therefore, the gun must have become unloaded between 2 and 5.[5] Our theory tells us nothing more than this. Consider the following acceptable models for *TI*′:

- $\mathcal{M}'_1$, where unload occurs at 2, the gun is unloaded at 3,4, and 5
- $\mathcal{M}'_2$, where unload occurs at 3, the gun is loaded at 3 and unloaded at 4 and 5
- $\mathcal{M}'_3$, where unload occurs at 4, the gun is loaded at 3 and 4, and unloaded at 5.

Intuitively, there does not seem to be a reason to prefer one of these models to the other. And in fact, our theory does not: $\mathcal{M}'_1$, $\mathcal{M}'_2$, and $\mathcal{M}'_3$ are equipreferable. Note, however, that both $\mathcal{M}'_1$ and $\mathcal{M}'_3$ are preferable to $\mathcal{M}'_4$, the model in which unload occurs at 2, load at 3, and unload at 4. $\mathcal{M}'_4$ is acceptable, but has superfluous actions. In fact, it can be shown that $\mathcal{M}'_1$, $\mathcal{M}'_2$, and $\mathcal{M}'_3$ are preferred models for *TI*′. All that *TI*′ can predict, then, is the disjunction:

$$\begin{aligned}\text{True(2,Occurs(unload))} \;\vee\; & \text{True(3,Occurs(unload))} \\ \vee\; & \text{True(4,Occurs(unload))}\end{aligned}$$

which is exactly what we wish.

4 Explanation

A theory of temporal reasoning that can handle both forward and backward projection properly is clearly a prerequisite for any theory of explanation. Now that we have developed such a theory, we present a theory of explanation.

Intuitively, the need to explain something arises when we are initially given some partial chronicle description accompanied by some theory, we make some projections, and then we subsequently discover these projections to be false. When we find out the true story, we feel a need to explain "*what went wrong*"—that is, why the original projections did not in fact hold true.

Formally, we can describe the situation as follows: Consider a theory instantiation $TI_1 = T \cup CD_1$, with $\cap_{\mathcal{M}^*(\mathrm{TI}_1)}$ equal to the set of facts projected by TI_1. Consider now a second theory instantiation $TI_2 = T \cup CD_2$, where $CD_2 \supset CD_1$. That is, TI_2 is TI_1 with a more fleshed out description of the chronicle. We say that there is a ***need for explanation of** TI_2 **relative to** TI_1* if there exists some fact $\kappa \in CD_2$ such that TI_1 does not project κ, *i.e.* if $(\exists \kappa \in CD_2)[\kappa \notin \cap_{\mathcal{M}^*(\mathrm{TI}_1)}]$. For any such κ, we say that κ ***must be explained*** relative to TI_1 and TI_2.

The need for explanation may be more or less pressing depending upon the particular situation. There are two cases to be distinguished:

Case I :

κ is not projected by TI_1, *i.e.* $\kappa \notin \cap_{\mathcal{M}^*(\mathrm{TI})}$. However κ is consistent with TI_1, *i.e.* $\kappa \in \cup_{\mathcal{M}^*(\mathrm{TI}_1)}$. That is, κ is true in some of the preferred models of TI_1, it just is not true in all of the preferred models. For example, consider $TI_1 = T \cup CD_1$, where T is the theory described in the previous section, and CD_1 = {True(1,loaded),True(2,¬loaded)}, and $TI_2 = T \cup CD_2$, where $CD_2 = CD_1 \cup$ {True(1,Occur(unload))}.

The set of preferred models for TI_1 contains models in which the gun becomes unloaded via an unload action, and models in which the gun becomes unloaded via a shoot action. Neither action is in the intersection of the preferred

[5] As we know, either an unload or a shoot will cause a gun to be unloaded. However, because we know that shooting will cause Fred to be dead, that dead persists forever, and that Fred is alive at 6, all acceptable models for *TI*′ must have an unload.

models, so neither action is projected by TI_1. TI_1 will only project that one of the actions must have occurred; *i.e.* the disjunct True(1,Occurs(shoot)) $\vee$ True(1,Occur(unload)).

The extra information in CD_2 does not contradict anything we know; it simply gives us a way of pruning the set of preferred models. Intuitively, an explanation in such a case should thus characterize the models that are pruned.

Case II :

κ is not projected by TI_1. In fact, κ is not even consistent with TI_1, *i.e.* $\kappa \notin \cup_{\mathcal{M}^*(TI_1)}$. In this case, it is in fact the case that $\neg\kappa \in \cap_{\mathcal{M}^*(\mathrm{TI}_1)}$, *i.e.*, TI_1 projects $\neg\kappa$.

Such a situation is in fact what we have in the Yale Shooting Scenario, if we find out, after predicting Fred's death, that he is indeed alive at time 6. This is the sort of situation that demonstrates the non-monotonicity of our logic, for TI_1 projects True(6,$\neg$alive), while $TI_2 \supset TI_1$ projects True(6,alive). Here the need for explanation is crucial; we must be able to explain why our early projection went awry.

Intuitively, an informal explanation of what went wrong in this case must contain the facts that an unload occurred and that the gun was thus unloaded at time 5. That is, an adequate explanation is an account of the facts leading up to the discrepancy in the chronicle description.

We formalize these intuitions as follows: Given TI_1, TI_2, and a set of facts Q which are unprojected by TI_1, we define an adequate explanation for the set of facts Q relative to TI_1 and TI_2 as the set difference between the projections of TI_2 and the projections of TI_1:

Definition: Let $\mathsf{Q} = \{\kappa \mid \kappa \in CD_2 \wedge \kappa \notin \cap_{\mathcal{M}^*(\mathrm{TI})}\}$

An adequate explanation for Q is given by $\cap_{\mathcal{M}^*(TI_2)} - \cap_{\mathcal{M}^*(TI_1)}$

As an example, let $TI_1 = T \cup CD_1$ be the description of the Yale Shooting Scenario (as in the previous section); let $TI_2 = T \cup CD_2$, where $CD_2 = CD_1 \cup \{\text{True(6,alive)}\}$. The explanation of True(6,alive) relative to TI_1 and TI_2 would include the facts that an unload occurred either at time 2 or time 3 or time 4, and that the gun was unloaded at time 5—precisely the account which we demand of an explanation.

5 Conclusions and Future Work

We have developed a theory of default temporal reasoning which allows us to perform temporal projection correctly. Central to our theory is the concept that models with the fewest possible unmotivated actions are preferred.

We have demonstrated that this theory handles both forward and backward temporal projection accurately. We have given an intuitive account of the ways in which the need for explanation arises, and have shown how we can define explanation in a natural way in terms of our theory of projection.

We are currently extending the work described in this paper in two directions. We are examining several different characterizations of the explanation process, and determining the relationships between these characterizations within our model. In addition, we are investigating the properties of a theory which minimizes unmotivated state changes, as opposed to unmotivated actions. Preliminary investigations suggest that such a theory would eliminate the need for both persistence rules and the principle of inertia.

Acknowledgements:

We'd like to thank Ernie Davis, Tom Dean, Vladimir Lifschitz, John McCarthy, and Yoav Shoham for ideas, advice, and many helpful discussions.

References

[Hanks and McDermott, 1986] Steven Hanks and Drew McDermott, "Default Reasoning, Nonmonotonic Logics, and the Frame Problem", *Proc. AAAI*, 1986.

[Haugh, 1987] Brian Haugh "Simple Causal Minimizations for Temporal Persistence and Projection", *Proc. AAAI*, 1987.

[Hayes, 1985] Patrick Hayes, "Naive Physics I: Ontology for Liquids", in J. Hobbs and R. Moore, editors, *Formal Theories of the Commonsense World*, Ablex 1985.

[Kautz, 1986] Henry Kautz, "The Logic of Persistence", *Proc. AAAI*, 1986.

[Lifschitz, 1986] Vladimir Lifschitz, "Pointwise Circumscription: Preliminary Report", *Proc. of AAAI*, 1986.

[Lifschitz, 1987] Vladimir Lifschitz, "Formal Theories of Action: Preliminary Report", *Proc. of IJCAI*, 1987.

[McCarthy, 1980] John McCarthy, "Circumscription—A Form of Nonmonotonic Reasoning", *Artificial Intelligence*, Vol. 13, 1980.

[McCarthy and Hayes, 1969] John McCarthy and Patrick Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence", In Donald Michie and Bernard Meltzer, editors, *Machine Intelligence*, Vol. 4, 1969.

[McDermott and Doyle, 1980] Drew McDermott and Jon Doyle, "Non-Monotonic Logic I", *Artificial Intelligence*, Vol. 13, 1980.

[McDermott, 1982] Drew McDermott, "A Temporal Logic for Reasoning about Processes and Plans", *Cognitive Science*, Vol. 6, 1982.

[McDermott, 1984] Drew McDermott, "The Proper Ontology for Time", Unpublished paper, 1984.

[McDermott 1987] Drew McDermott, "AI, Logic, and the Frame Problem", *Proc. The Frame Problem in Artificial Intelligence*, 1987.

[Reiter, 1980] Ray Reiter, "A Logic for Default Reasoning", *Artificial Intelligence*, Vol. 13, 1980.

[Shoham, 1986] Yoav Shoham, "Chronological Ignorance: Time, Nonmonotonicity, Necessity, and Causal Theories", *Proc. AAAI*, 1986.

[Shoham, 1987] Yoav Shoham, "Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence", Phd Thesis, Tech. Report 507, Yale Univ. 1987.