Motivated Action Theory: A Formal Theory of Causal Reasoning

Lynn Andrea Stein
Leora Morgenstern
Brown University Department of Computer Science
Providence, RI 02912, U.S.A.

Technical Report No. CS-89-12, March 1989

Motivated Action Theory: A Formal Theory of Causal Reasoning

Lynn Andrea Stein and Leora Morgenstern *

Department of Computer Science
Brown University
Box 1910, Providence, RI 02912
{las,lm}%cs.brown.edu@relay.cs.net

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.

*A shorter version of this paper appears in *AAAI 88* as "Why things go wrong: a formal theory of causal reasoning." The authors are supported by an IBM graduate fellowship and grant no. NSF–IRI–8801253, respectively.

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, 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 on the idea 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 Yale shooting problem, 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 with respect to 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 him 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]).

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

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 preorder; 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]This point was noted by Kautz when he first presented his solution to the Yale shooting problem.

2.2.2 Circumscribing over causes

The temporal projection problem is actually a dual one. We must reason that unexpected events don't happen, and that events that *do* occur don't cause states to change in unexepcted ways. Much of the work on the Yale shooting problem has focused on the first of these problems. The solutions proposed by [Lifschitz, 1987] and [Haugh, 1987] address the second.

Lifschitz's and Haugh's solutions are not based on forward reasoning strategies. Rather, they work by minimizing the changes caused by an action. They introduce a predicate causes(act,f,v), where action act causes fluent f to take on value v. Circumscribing over causes addresses the second temporal projection problem: actions don't have strange effects.[3]

Lifschitz solves the first temporal projection problem by defining it away: his theory works in the situation calculus, where every state is the result of performing a specific action in a specific (previous) state. In this framework, he solves the frame problem by circumscribing over causes: 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, it is highly problematic. It depends on the situation calculus to provide all and exactly those actions that do occur. Consider what would happen in a world in which concurrent (or uncertain) 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. This cannot in fact happen in Lifschitz's formulation because in the rigid *STRIPS* framework, 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.

[3] Lifschitz and Haugh also introduce the precond predicate; they use precond to solve the qualification problem, which we don't discuss here.

Lifschitz's solution thus works only in frameworks where all the events in a chronicle are known. In these cases, circumscribing the **causes** 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.

Haugh's solution shares the **causes** predicate with Lifschitz's, but also incorporates an explicit solution to the problem of preventing unwanted actions. Rather than minimizing the extent of **causes**, he circumscribes over a predicate **potential-causes**, which is the conjunction of **causes** and **occurs**.[4] Since **unload** causes some effects, *if* it occurred, it would be a **potential-cause**. We can't minimize **causes(unload,loaded,false)**—it is in our axiomatization. So instead we minimize **occurs(unload...)**, and the **unload** never happens.

This is actually quite effective in many situation. However, Haugh's theory has some strange results with respect to disjunction. If, for example, we know that, while we were out of the room, either nothing happens (**wait**) or someone **unloads** the gun. Haugh's theory of potential causes predicts that the **wait** occurs, and the gun remains loaded. Unlike **unload**, **wait** has no effects, so there are no **causes** axioms on **wait**. This means that even if **wait** occurs, it won't be a **potential-cause**. **Unload**, as we've seen above, is a **potential-cause** whenever it **occurs**.

These difficulties in Haugh's theory arise from a confusion between *action* and *state change*. There are many actions without obvious state changes—McDermott, e.g., suggests "run around the track three times" [McDermott, 1982, p. 109]. Since Haugh is concerned with the conjunction of **causes** and **occurs**, he is really only interested in minimizing the occurrence of actions with effects. In his framework, "run around the track three times" can occur arbitrarily often, just as **unload** can occur arbitrarily often in Lifschitz's framework. Since "run around the track three times" has no effects, it is never a **potential-cause**. Minimizing potential causes can never eliminate a "run around the track three times" event.

Haugh's theory suffers from a second, though perhaps less disconcerting, anomaly. Since the theory only minimizes **causes** for actions that actually **occur**, actions that never occur can have arbitrary effects. For example, patting my stomach and rubbing my head could cause the world to blow up, provided I never actually *do* pat my stomach and rub my head. Of course, if I ever did succeed in patting my stomach and rubbing my head, this bizarre effect would go away, but it is somewhat strange to allow a conclusion like **causes(pat-and-rub,True,blow-up(world))**.

[4] Haugh actually uses result rather than occurs, but the effects are essentially the same. He also presents two solutions: *potential causes* and *determined causes*. Haugh's theory of determined causes suffers from anomalous results similar to, but more severe than, those described here for potential causes.

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 Notation

We begin by formally describing the concepts of a theory and a chronicle description. We work in a first order logic $\mathcal{L}$, augmented by a simple temporal logic. Terms 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.

We will simply write Occurs(j,act) to stand for terms of the form of the True(j,Occurs(act)). A term of the form Occurs(t,act) is called an *occurrence term.* A term of the form ¬Occurs(t,act) is a *negative occurrence term.* Given a set A of occurrence terms, we define its complement, $\overline{\mathtt{A}}$, to be the set of negative occurrence terms $\{\neg\text{Occurs(t,act)} \mid \text{Occurs(t,act)} \notin \mathtt{A}\}$. A term of the form True(t,fluent), or of the form True(t,¬fluent), where fluent is not of the form (¬)Occurs(act) is a *state term*, and such a fluent is a *state.*[5]

A *theory*, T, and a *chronicle description*, CD, are sets of sentences of $\mathcal{L}$. 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

$$\alpha \wedge \beta \supset \gamma$$

where:

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

β is a conjunction of terms (including no positive occurrence terms) giving the preconditions of the action, and

[5] Unlike actions and their negations, which can generally be distinguished ("crossing the street" is an action; "not crossing the street" is not), it is often impossible to tell which of a state and its negation is "positive" (e.g. "light" vs. "dark").

γ describes the results of the action.

Note that γ can include occurrence terms. We can thus express causal chains of action. If $\alpha \wedge \beta \supset \gamma$, we will say that $\gamma \in \mathsf{Results}(\alpha,\beta)$.

A persistence rule is of the form

$$\mathsf{True(j,f)} \wedge \beta \supset \mathsf{True(j+1,f)}$$

where f is a state, and β includes no positive occurrence terms (Typically, β will be a conjunction of negative occurrence terms). 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):

$$\begin{aligned}\mathsf{True(j,f)} \wedge (\forall \mathsf{act}.(\mathsf{f} \in \mathsf{Results(act,preconds)} \\ \supset (\neg \mathsf{act} \vee \neg \mathsf{preconds})) \\ \supset \mathsf{True(j+1,f)}\end{aligned}$$

We have hand coded the persistence rules for this simple case, although it is not necessary to do so. They can 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 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 criteria on models:[6] in particular, preferring models in which the fewest possible extraneous actions occur. Typically, we will not be given 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: model theory

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

[6] In section 3.3, we give a proof tehoretic version of motivated action theory, but the axioms remain monotonic. Here, we introduce a sort of "syntactic circumscription" or preference over sets of sentences, making the monotonic theory non-monotonic through the introduction of a new rule of inference.

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 $\mathtt{TI} = \mathtt{T} \cup \mathtt{CD}$, we are thus interested in determining the preferred models for TI. $\mathcal{M}(\mathtt{TI})$ denotes a model for TI: *i.e.*, $(\forall \varphi \in \mathtt{TI})[\mathcal{M}(\mathtt{TI}) \models \varphi]$. 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 $\mathtt{TI} = \mathtt{T} \cup \mathtt{CD}$, we say that a statement φ is *motivated in* $\mathcal{M}(\mathtt{TI})$ if it is strongly motivated in $\mathcal{M}(\mathtt{TI})$ or weakly motivated in $\mathcal{M}(\mathtt{TI})$.

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

A statement φ is *weakly* motivated in $\mathcal{M}(\mathtt{TI})$ if there exists in TI a causal rule of the form $\alpha \wedge \beta \supset \varphi$, α is (strongly or weakly) motivated in $\mathcal{M}(\mathtt{TI})$, and $\mathcal{M}(\mathtt{TI}) \models \beta$.[7]

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)). $\mathcal{M}_i(\mathtt{TI}) \trianglelefteq \mathcal{M}_j(\mathtt{TI})$ ($\mathcal{M}_i$ is *preferable* to $\mathcal{M}_j$) if $(\forall \varphi)[$ if $\mathcal{M}_i(\mathtt{TI}) \models \varphi \wedge \mathcal{M}_j(\mathtt{TI}) \not\models \varphi$ then φ is motivated in $\mathcal{M}_i(\mathtt{TI})]$.

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

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

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

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

[7] In the original formulation of motivated action theory, this definition included persistence rules. However, motivation is only relevant for actions—we minimize unmotivated actions—making persistence rules, which would allow motivation only of *state*, irrelevant.

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

We define the following sets:

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

$\cup_{\mathcal{M}^*} = \{\varphi \mid \exists \mathcal{M} \in \mathcal{M}^*(\mathtt{TI}), \mathcal{M} \models \varphi\}$—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 $\cap_{\mathcal{M}^*(\mathtt{TI})}$. In this case, we say that TI *projects* φ.

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

Case III: φ not in $\cup_{\mathcal{M}^*(\mathtt{TI})}$. 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 Motivated actions: proof theory

The proof theory for motivated actions is based on the construction of sets of sentences analogous to models. We then transform the preference criterion defined on models in the previous section to one defined on these sets of sentences; the theorems of motivated action theory are exactly those sentences contained in the most-preferred set.

Definition: An *occurence kernel* is a pair <A,B>, where A is a set of occurence terms and B is a set of state terms. An occurence kernel <A,B> is *acceptable* for a theory instantiation TI if $\mathtt{TI} \cup \mathtt{A} \cup \mathtt{B} \cup \overline{\mathtt{A}}$ is consistent. In this case, we write $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}}$for $\mathtt{TI} \cup \mathtt{A} \cup \mathtt{B} \cup \overline{\mathtt{A}}$. We say that $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}}$ *supports* a statement φ if $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \varphi$.

An occurrence kernel thus determines the complete set of actions that do (A) and don't ($\overline{\mathtt{A}}$) occur. If B provides a value for every state at every time, then the "world" of <A,B> is completely determined—actions by A and $\overline{\mathtt{A}}$, and state by B. However, we do not in general need such a complete B. If the value of a state σ is determined once, its value at every later time point can be determined by examining the actions that do or don't occur, affecting it. But all actions are fully determined, since $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}}$ includes $\overline{\mathtt{A}}$. The only further requirement is that if an action $\varphi \in \mathtt{A}$ might change σ, (and, since $\varphi \in \mathtt{A}$, we

know that it occurs), the truth value of all of the preconditions to a must be determined by $\mathsf{time}(\varphi)$

Formally, a state σ is *determined* at time t (in $<\mathtt{A},\mathtt{B}>$) if

1. $\mathsf{True}(\mathsf{t},(\neg)\sigma) \in \mathtt{B}$, or
2. $\varphi =\in \mathtt{A}$, and $(\neg)\sigma \in \mathsf{Results}(\varphi,\mathsf{preconds})$ and $\forall\pi \in \mathsf{preconds}, <\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \pi$, or
3. $(\exists \mathsf{t}' < \mathsf{t})[\sigma$ determined at $\mathsf{t}']$, and
   for every occurrence term $\varphi \in \mathtt{A}$, if $(\mathsf{t}' \leq \mathsf{time}(\varphi) < \mathsf{t}$ and $\sigma \in \mathsf{Results}(\varphi,\mathsf{preconds}))$ then $(\forall\pi = \mathsf{True}(\mathsf{t}'',\mathsf{f}) \in \mathsf{preconds})$ [f is determined at t''].

All actions are determined at all times.

Definition: An occurence kernel, $<\mathtt{A},\mathtt{B}>$, is *total* for TI if $\forall\varphi \in \mathtt{A}$.(if $\sigma \in \mathsf{Results}(\varphi,\mathsf{preconds})$ then σ is determined at $\mathsf{time}(\varphi)+1$), and if $\alpha \wedge \beta \supset \varphi \in \mathtt{T}$ and $\varphi \in \mathtt{A}$, then β is determined.

Total occurrence kernels determine the results of all actions; in this sense, they correspond to sets of models, or limited world views. We will assume all occurrence kernels below to be total. We take $\mathcal{OC}(\mathtt{TI})$ to be the set of occurrence kernels $<\mathtt{A},\mathtt{B}>$ that are both acceptable and total for TI.

We now define the syntactic equivalent of motivation, the predicate $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$ recursively in terms of the first-order consequences of $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}}$.

Definition: $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$ if

1. $\mathtt{TI} \supset \varphi$, or
2. there exists in TI a causal rule of the form $\alpha \wedge \beta \rightarrow \varphi$; $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\alpha)$; and $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \beta$.

mot induces a preorder on occurrence kernels. If $<\mathtt{A},\mathtt{B}>$ and $<\mathtt{A}',\mathtt{B}'>$ are occurrence kernels in $\mathcal{OC}(\mathtt{TI})$, we say that $<\mathtt{A},\mathtt{B}>$ is preferred to $<\mathtt{A}',\mathtt{B}'>$ ($<\mathtt{A},\mathtt{B}> \preceq <\mathtt{A}',\mathtt{B}'>$) if for every action φ, if $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \varphi$ and $<\mathtt{A}',\mathtt{B}'>_{\mathtt{TI}} \nvdash \varphi$, then $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$. As with models, we call minimal elements under this ordering *preferred,* and call the set of these preferred occurrence kernels $\mathcal{OC}^*(\mathtt{TI})$. We define $\cup_{\mathcal{OC}^*(\mathtt{TI})}$ and $\cap_{\mathcal{OC}^*(\mathtt{TI})}$ to be the union and intersection of statements in preferred occurrence kernels, respectively.

Below, we show that this definition of motivation is both sound and complete with respect to the semantic notion of motivation. First, we define a particular mapping from models to occurrence kernels, and show that it preserves motivation. Similarly, the mapping from occurrence kernels to models is motivation-preserving. Proofs of these lemmas are given in appendix B.

Definition: $\mathcal{OC}_{\mathcal{M},\mathtt{TI}}$, the occurrence kernel of model $\mathcal{M}$ for theory instantiation TI, is the occurrence kernel <A, B> given by

$$\begin{aligned}\mathtt{A} &= \{\alpha \mid \alpha \text{ an occurrence term and } \mathcal{M} \models \alpha\}\\ \mathtt{B} &= \{\beta \mid \beta \text{ a state term and either } \exists \alpha \in \mathtt{A}, (\neg)\beta \in \mathsf{Results}(\alpha,\mathsf{preconds})\\ &\qquad\qquad \text{or } \alpha \wedge (\neg)\beta \wedge \beta' \supset \varphi \in \mathtt{T} \text{ and } \varphi \in \mathtt{A},\\ &\qquad \text{and } \mathcal{M} \models \beta\}\end{aligned}$$

Lemma 1 *Let $\mathcal{M}$ be a model for a theory instantiation* TI, *and let* $<\mathtt{A},\mathtt{B}> = \mathcal{OC}_{\mathcal{M},\mathtt{TI}}$. *Then ($\mathcal{OC}_{\mathcal{M},\mathtt{TI}}$ is acceptable and total for* TI, *and for every action φ, $\mathcal{M} \models \varphi$ iff* $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash$ *varphi.*

Lemma 2 *Let $\mathcal{M}$ be a model for a theory instantiation* TI, *and let* $<\mathtt{A},\mathtt{B}> = \mathcal{OC}_{\mathcal{M},\mathtt{TI}}$. *If $\mathcal{M} \models \varphi$ (or* $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \varphi$*), then* $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$ *iff φ is motivated in $\mathcal{M}$.*

Lemma 3 *Let* $<\mathtt{A},\mathtt{B}>$ *be an occurrence kernel acceptable and total for a theory instantiation* TI. *Then if $\mathcal{M}$ is a model for* $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}}$, $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$ *iff φ is motivated in $\mathcal{M}$.*

Theorem 1 (Soundness and completeness) *Let* TI *be a theory instantiation, with $\mathcal{M}^\star$ the set of preferred models for* TI, *and $\mathcal{OC}^\star$ the set of preferred occurrence kernels for* TI. *Then $\varphi \in \cup_{\mathcal{M}^\star(\mathtt{TI})}$ iff $\varphi \in \cup_{\mathcal{OC}^\star(\mathtt{TI})}$.*
Proof.

$\varphi \in \cup_{\mathcal{M}^\star(\mathtt{TI})} \Rightarrow \varphi \in \cup_{\mathcal{OC}^\star(\mathtt{TI})}$.

Assume, for some φ, that $\varphi \in \cup_{\mathcal{M}^\star(\mathtt{TI})}$. Then there is some model $\mathcal{M} \in \mathcal{M}^\star(\mathtt{TI})$ such that $\mathcal{M} \models \varphi$. Let $<\mathtt{A},\mathtt{B}> = \mathcal{OC}_{\mathcal{M},\mathtt{TI}}$ be the occurrence kernel of that model. We claim that $<\mathtt{A},\mathtt{B}> \in \mathcal{OC}^\star(\mathtt{TI})$—$<\mathtt{A},\mathtt{B}>$ is a preferred occurrence kernel—and so $\varphi \in \cup_{\mathcal{OC}^\star(\mathtt{TI})}$.

Assume that $<\mathtt{A},\mathtt{B}>$ is not preferred. Then there is some occurrence kernel $<\mathtt{A}',\mathtt{B}'>_{\mathtt{TI}}$ such that $<\mathtt{A}',\mathtt{B}'>_{\mathtt{TI}} \preceq <\mathtt{A},\mathtt{B}>_{\mathtt{TI}}$, but $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \not\preceq <\mathtt{A}',\mathtt{B}'>_{\mathtt{TI}}$. That is, for every action ω, if $<\mathtt{A}',\mathtt{B}'>_{\mathtt{TI}} \vdash \omega$ and $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \nvdash \omega$, then $\mathsf{mot}(\mathtt{A}',\mathtt{B}',\mathtt{TI},\omega)$; but for some action ω', $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \omega'$, $<\mathtt{A}',\mathtt{B}'>_{\mathtt{TI}} \nvdash \omega'$, and $\neg\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\omega')$. But $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\omega)$ iff ω is motivated in $\mathcal{M}$, and—given $\mathcal{M}'$, some model for $<\mathtt{A}',\mathtt{B}'>_{\mathtt{TI}}$—$\mathsf{mot}(\mathtt{A}',\mathtt{B}',\mathtt{TI},\omega)$ iff ω is motivated in $\mathcal{M}'$ (by Lemmas 2 and 3), and occurrence terms are derivable from occurence kernels iff they are entailed by the corresponding models (by Lemma 1). Hence, $\mathcal{M}' \trianglelefteq \mathcal{M}$, but $\mathcal{M} \not\trianglelefteq \mathcal{M}'$, and so $\mathcal{M}$ is not maximally preferred. Contradiction.

$\varphi \in \cup_{\mathcal{OC}^\star(\mathtt{TI})} \Rightarrow \varphi \in \cup_{\mathcal{M}^\star(\mathtt{TI})}$.

Assume, for some φ, that $\varphi \in \cup_{\mathcal{OC}^\star(\mathtt{TI})}$. Then there is some occurrence kernel $<\mathtt{A},\mathtt{B}> \in \mathcal{OC}^\star(\mathtt{TI})$ such that $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \varphi$. Let $\mathcal{M}$ be a model for $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}}$. We claim that $\mathcal{M} \in \mathcal{M}^\star(\mathtt{TI})$—$\mathcal{M}$ is a preferred model for TI—and so $\varphi \in \cup_{\mathcal{M}^\star(\mathtt{TI})}$.

Assume that $\mathcal{M}$ is not preferred. Then there is some model $\mathcal{M}'$ such that $\mathcal{M}' \trianglelefteq \mathcal{M}$, but $\mathcal{M} \not\trianglelefteq \mathcal{M}'$. That is, for every action ω, if $\mathcal{M}' \models \omega$ and $\mathcal{M} \not\models \omega$, then ω is motivated in $\mathcal{M}'$; but for some action ω', $\mathcal{M} \models \omega'$, $\mathcal{M}' \not\models \omega'$, and ω' is not motivated in $\mathcal{M}$. But ω is motivated in $\mathcal{M}$ iff $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\omega)$, and if $<\mathtt{A}',\mathtt{B}'> =$, ω is motivated in $\mathcal{M}'$ iff $\mathsf{mot}(\mathtt{A}',\mathtt{B}',\mathtt{TI},\omega)$(by Lemmas 2 and 3), and occurrence terms are derivable from occurence kernels iff they are entailed by the corresponding models (by Lemma 1). Hence, $<\mathtt{A}',\mathtt{B}'> \preceq <\mathtt{A},\mathtt{B}>$, but $<\mathtt{A},\mathtt{B}> \not\preceq <\mathtt{A}',\mathtt{B}'>$, and so $<\mathtt{A},\mathtt{B}>$ is not maximally preferred. Contradiction.

Corollary 1 *Let* TI, $\mathcal{M}^\star(\mathtt{TI})$, *and* $\mathcal{OC}^\star(\mathtt{TI})$ *be as above. Then if* φ *is a (positive or negative) occurrence term, or if there is a causal rule of the form* $\alpha \wedge \beta \supset \varphi$ *in* T, *then* $\varphi \in \cap_{\mathcal{M}^\star(\mathtt{TI})}$ *iff* $\varphi \in \cap_{\mathcal{OC}^\star(\mathtt{TI})}$.

Proof.

$\varphi \in \cap_{\mathcal{OC}^\star(\mathtt{TI})}$ means that for every occurrence kernel $<\mathtt{A},\mathtt{B}>$ in $\mathcal{OC}^\star(\mathtt{TI})$, $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \varphi$. But then certainly not $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \neg\varphi$. So $\neg\varphi \notin \cup_{\mathcal{OC}^\star(\mathtt{TI})}$, and hence $\neg\varphi \notin \cup_{\mathcal{M}^\star(\mathtt{TI})}$. Since models entail p or $\neg p$, for every p, $\neg\neg\varphi = \varphi \in \cap_{\mathcal{M}^\star(\mathtt{TI})}$. Similarly, $\varphi \in \cap_{\mathcal{M}^\star(\mathtt{TI})}$ means that every model $\mathcal{M}$ in $\mathcal{M}^\star(\mathtt{TI})$ entails φ, so $\neg\varphi \notin \cup_{\mathcal{M}^\star(\mathtt{TI})}$ and hence $\notin \cup_{\mathcal{OC}^\star(\mathtt{TI})}$. φ an occurrence term, or $\alpha \wedge \beta \supset \varphi$ in T, means that every occurrence kernel in $\mathcal{OC}^\star(\mathcal{M},\mathtt{TI})$ is definite for φ, so $\neg\varphi \notin \cup_{\mathcal{OC}^\star(\mathtt{TI})}$ implies $\neg\neg\varphi = \varphi \in \cap_{\mathcal{OC}^\star(\mathtt{TI})}$.

3.4 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:

$$\mathsf{True}(1,\mathsf{alive})$$
$$\mathsf{True}(1,\mathsf{Occurs}(\mathsf{load}))$$
$$\mathsf{True}(5,\mathsf{shoot})$$

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

T: Causal Rules:

$$\mathsf{True}(j,\mathsf{Occurs}(\mathsf{load})) \supset \mathsf{True}(j{+}1,\mathsf{loaded})$$

$$\text{True(j,Occurs(shoot))} \wedge \text{True(j,loaded)} \supset \text{True(j} + 1, \neg\text{alive)}$$
$$\text{True(j,shoot)} \supset \text{True(j} + 1, \neg\text{loaded)}$$
$$\text{True(j,unload)} \supset \text{True(j} + 1, \neg\text{loaded)}$$

Persistence Rules:

$$\text{True(j,alive)} \wedge (\text{True(j,} \neg\text{Occurs(shoot))} \vee \text{True(j,} \neg\text{loaded))} \supset \text{True(j+1,alive)}$$
$$\text{True(j,loaded)} \wedge \text{True(j,} \neg\text{Occurs(shoot))} \wedge \text{True(j,} \neg\text{Occurs(unload))} \supset \text{True(j+1,loaded)}$$

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 we have the causal rules

$$\text{True(j,Occurs(passAB))} \wedge \text{True(j,on)} \supset \text{True(j+1,Occurs(passBC))}$$
$$\text{True(j,Occurs(passBC))} \wedge \text{True(j,on)} \supset \text{True(j+1,Occurs(passCD))}$$
$$\text{True(j,Occurs(turnoff))} \supset \text{True(j+1,off)}$$

the persistence rule

$$\text{True(j,on)} \wedge \text{True(j,} \neg\text{Occurs(turnoff))} \supset \text{True(j+1,on)}$$

Assume the chronicle description contains: True(1,on) and True(1,Occurs(passAB)). Then, using our preference criterion, the theory instantiation projects

True(1,Occurs(passAB))
True(2,Occurs(passBC))
True(3,Occurs(passCD))

Minimizing actions would yield: True(1,Occurs(passAB)) and True(1,Occurs(turnoff)). The non-equivalence of the two criteria will hold in any theory with causal chains of events. Thus our criterion is not equivalent to circumscribing over the Occurs predicate.

3.5 Backward projection

We now show that our theory handles backward projection properly. As an example, consider TI′, where TI′ = TI ∪ {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.[8] 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:

True(2,Occurs(unload)) ∨ True(3,Occurs(unload)) ∨ True(4,Occurs(unload))

which is exactly what we wish.

[8] 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.

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 $\mathtt{TI}_1 = \mathtt{T} \cup \mathtt{CD}_1$, with $\cap_{\mathcal{M}^*(\mathtt{TI}_1)}$ equal to the set of facts projected by $\mathtt{TI}_1$. Consider now a second theory instantiation $\mathtt{TI}_2 = \mathtt{T} \cup \mathtt{CD}_2$, where $\mathtt{CD}_2 \supset \mathtt{CD}_1$. That is, $\mathtt{TI}_2$ is $\mathtt{TI}_1$ with a more fleshed out description of the chronicle. We say that there is a *a need for explanation of* $\mathtt{TI}_2$ *relative to* $\mathtt{TI}_1$ if there exists some fact $\kappa \in \mathtt{CD}_2$ such that $\mathtt{TI}_1$ does not project κ, *i.e.* if $(\exists \kappa \in \mathtt{CD}_2)[\kappa \ni \cap_{\mathcal{M}^*(\mathtt{TI}_1)}]$. For any such κ, we say that κ *must be explained* relative to $\mathtt{TI}_1$ and $\mathtt{TI}_2$.

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 $\mathtt{TI}_1$, *i.e.* $\kappa \ni \cap_{\mathcal{M}^*(\mathtt{TI})}$. However κ is consistent with $\mathtt{TI}_1$, *i.e.* $\kappa \in \cup_{\mathcal{M}^*(\mathtt{TI}_1)}$. That is, κ is true in some of the preferred models of $\mathtt{TI}_1$, it just is not true in all of the preferred models. For example, consider $\mathtt{TI}_1 = \mathtt{T} \cup \mathtt{CD}_1$, where T is the theory described in the previous section, and $\mathtt{CD}_1 = \{\mathsf{True}(1,\mathsf{loaded}), \mathsf{True}(2,\neg\mathsf{loaded})\}$, and $\mathtt{TI}_2 = \mathtt{T} \cup \mathtt{CD}_2$, where $\mathtt{CD}_2 = \mathtt{CD}_1 \cup \{\mathsf{True}(1,\mathsf{Occur}(\mathsf{unload}))\}$.

The set of preferred models for $\mathtt{TI}_1$ 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 models, so neither action is projected by $\mathtt{TI}_1$. $\mathtt{TI}_1$ will only project that one of the actions must have occurred; *i.e.* the disjunct $\mathsf{True}(1,\mathsf{Occurs}(\mathsf{shoot})) \vee \mathsf{True}(1,\mathsf{Occur}(\mathsf{unload}))$.

The extra information in $\mathtt{CD}_2$ 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 $\mathtt{TI}_1$. In fact, κ is not even consistent with $\mathtt{TI}_1$, *i.e.* $\kappa \ni \cup_{\mathcal{M}^*(\mathtt{TI}_1)}$. In this case, it is in fact the case that $\neg\kappa \in \cap_{\mathcal{M}^*(\mathtt{TI}_1)}$, *i.e.*, $\mathtt{TI}_1$ 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 $\mathtt{TI}_1$ projects True(6,¬alive), while $\mathtt{TI}_2 \supset \mathtt{TI}_1$ 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 $\mathtt{TI}_1$, $\mathtt{TI}_2$, and a set of facts Q which are unprojected by $\mathtt{TI}_1$, we define an adequate explanation for the set of facts Q relative to $\mathtt{TI}_1$ and $\mathtt{TI}_2$ as the set difference between the projections of $\mathtt{TI}_2$ and the projections of $\mathtt{TI}_1$:

Definition: Let $\mathsf{Q} = \{\kappa \mid \kappa \in \mathtt{CD}_2 \wedge \kappa \ni \cap_{\mathcal{M}^*(\mathtt{TI})}\}$

An adequate explanation for Q is given by $\cap_{\mathcal{M}^*(\mathtt{TI}_2)} - \cap_{\mathcal{M}^*(\mathtt{TI}_2)}$

As an example, let $\mathtt{TI}_1 = \mathtt{T} \cup \mathtt{CD}_1$ be the description of the Yale shooting scenario (as in the previous section); let $\mathtt{TI}_2 = \mathtt{T} \cup \mathtt{CD}_2$, where $\mathtt{CD}_2 = \mathtt{CD}_1 \cup \{\mathsf{True(6,alive)}\}$. The explanation of True6alive relative to $\mathtt{TI}_1$ and $\mathtt{TI}_2$ 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.

Note that, due to our preference criterion, explanations in this theory are minimal in the number of unmotivated actions that they posit. The theory thus lends itself to the goal of finding the simplest possible explanation for an unexpected outcome.

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 Drew McDermott, Ernie Davis and Tom Dean 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. In *AAAI*, 1986.

[Haugh, 1987] Brian Haugh. Simple causal minimizations for temporal persistence and projection. In *AAAI*, 1987.

[Hayes, 1985] Patrick Hayes. Naive physics i: Ontology for liquids. In Jerry Hobbs and Robert Moore, editors, *Formal Theories of the Commonsense World.* Ablex, 1985.

[Kautz, 1986] Henry Kautz. The logic of persistence. In *AAAI*, 1986.

[Lifschitz, 1986] Vladimir Lifschitz. Pointwise circumscription: Preliminary report. In *AAAI*, 1986.

[Lifschitz, 1987] Vladimir Lifschitz. Formal theories of action: Preliminary report. In *IJCAI*, 1987.

[McCarthy and Hayes, 1969] John McCarthy and Patrick Hayes. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4, 1969.

[McCarthy, 1980] John McCarthy. Circumscription—a form of nonmonotonic reasoning. *Artificial Intelligence*, 13, 1980.

[McDermott and Doyle, 1980] Drew McDermott and Jon Doyle. Non-monotonic logic I. *Artificial Intelligence*, 13, 1980.

[McDermott, 1982] Drew McDermott. A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6, 1982.

[McDermott, 1984] Drew McDermott. The proper ontology for time. Unpublished paper, 1984.

[McDermott, 1987] Drew McDermott. AI, logic, and the frame problem. In Frank Brown, editor, *The Frame Problem in AI.* Morgan Kauffman, 1987.

[Reiter, 1980] Ray Reiter. A logic for default reasoning. *Artificial Intelligence*, 13, 1980.

[Shoham, 1986] Yoav Shoham. Chronological ignorance: Time, nonmonotonicity, necessity, and causal theories. In *AAAI*, 1986.

[Shoham, 1987] Yoav Shoham. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence.* PhD thesis, Yale University Department of Computer Science, 1987. Yale University Department of Computer Science Research Report 507.

A Complex Expressions in CD

Throughout this paper, we have treated CD as though it might contains only simple terms without connectives. In fact, the mechanisms we've described can handle more, but we need to make a few adjustments in order to deal properly with all of the possibilities that seem reasonable. Since the truth values of state terms are generally allowed to vary in the nonmonotonic part of MAT, we are really only concerned with the handling of occurence terms. The definitions given here work as well if there are state and/or negative occurence terms involved; since their main force is to elaborate the definition of motivation; however, their effect for state/negative occurence terms is negligible.[9]

If a *conjunction* of occurence terms, $\varphi_1 \wedge \ldots \wedge \varphi_n$, is in CD, we can treat this as though the individual occurence terms, φ_i, appear in CD. The definition of motivation already deals with these actions properly: $\mathtt{TI} \vdash \varphi_i$, so $\mathsf{mot}(\mathtt{A}, \mathtt{B}, \mathtt{TI}, \varphi_i)$, $1 \leq i \leq n$; similarly, φ_i is motivated in all models $\mathcal{M}_{\mathtt{TI}}$.

If a *disjunction* of occurence terms, $\varphi_1 \vee \cdots \vee \varphi_n$, is in CD, we certainly cannot just break up the disjunction and add the individual occurence terms to CD. However, while this is obviously incorrect for assertion, it turns out to be precisely the right thing to do for motivation. For disjunctions, then, we add the definition of semi-motivation:

Definition: An occurence term φ is *semi-motivated in* TI if there is a disjunction $\varphi \vee \psi \vee \cdots \vee \psi_n$ of occurence terms in CD.

An occurence term φ is *semi-motivated in* $\mathcal{M}_{\mathtt{TI}}$ if φ is semi-motivated in TI, and $\mathcal{M}_{\mathtt{TI}} \models \varphi$.

Further, φ is motivated in $\mathcal{M}_{\mathtt{TI}}$ if φ is strongly motivated in $\mathcal{M}_{\mathtt{TI}}$, or φ is weakly motivated in $\mathcal{M}_{\mathtt{TI}}$, or φ is semi-motivated in $\mathcal{M}_{\mathtt{TI}}$. In the definition of weak motivation, α may be semi-motivated.

Syntactically, this corresponds to adding condition $1'$ to the definition of motivated:

$1.'$ CD contains a disjunction of occurence terms, $\varphi \vee \psi \vee \cdots \vee \psi_n$, or

Quantification can be handled in this framework by treating existential and universal quantifiers as infinite disjunctions and conjunctions, respectively.

[9]The authors are indebted to Matt Ginsberg for pointing out the difficulties which arise with disjunctions in CD.

B Proofs

Lemma 1 *Let $\mathcal{M}$ be a model for a theory instantiation* TI*, and let* $<\mathtt{A},\mathtt{B}> = \mathcal{OC}_{\mathcal{M},\mathtt{TI}}$. *Then ($\mathcal{OC}_{\mathcal{M},\mathtt{TI}}$ is acceptable and total for* TI*, and for every action φ, $\mathcal{M} \models \varphi$ iff* $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash$ *varphi.*
Proof.

$<\mathtt{A},\mathtt{B}>_{\mathtt{TI}}$ is acceptable:

$\mathcal{M} \models \mathtt{TI}$; $\mathcal{M} \models \mathtt{A},\mathtt{B}$; $\mathcal{M} \models \overline{\mathtt{A}}$ (since any action α such that $\mathcal{M} \models \alpha$ is in A, and $\mathcal{M} \not\models \alpha \Rightarrow \mathcal{M} \models \neg\alpha$. So $\mathcal{M} \models \mathtt{TI} \cup \mathtt{A} \cup \mathtt{B} \cup \overline{\mathtt{A}} = <\mathtt{A},\mathtt{B}>_{\mathtt{TI}}$, and $\mathtt{TI} \cup \mathtt{A} \cup \mathtt{B} \cup \overline{\mathtt{A}}$ is consistent.

$<\mathtt{A},\mathtt{B}>_{\mathtt{TI}}$ is definite:

This follows from the definitions of B and *definite.*

For every action φ, $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \varphi$ iff $\mathcal{M} \models \varphi$:

This follows directly from the definition of A.

Lemma 2 *Let $\mathcal{M}$ be a model for a theory instantiation* TI*, and let* $<\mathtt{A},\mathtt{B}> = \mathcal{OC}_{\mathcal{M},\mathtt{TI}}$. *If $\mathcal{M} \models \varphi$ (or* $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \varphi$*), then* $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$ *iff φ is motivated in $\mathcal{M}$.*
Proof.

For every action φ, if $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \varphi$ and $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$ then φ is motivated in $\mathcal{M}$:

If $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$, then either $\mathtt{TI} \supset \varphi$, and so φ is in all models of TI, or there is a causal rule of the form $\alpha \wedge \beta \supset \varphi$ in T, $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\alpha)$, and $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \supset \beta$. But then $\mathsf{time}(\alpha) < \mathsf{time}(\varphi)$, so we can show recursively that α is motivated in $\mathcal{M}$; further, $\mathtt{TI} \supset \beta \Rightarrow \mathcal{M} \models \beta$, since $\mathcal{M} \models \mathtt{TI}$.

For every action φ, if $\mathcal{M} \models \varphi$ and φ is motivated in $\mathcal{M}$ then $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$:

If φ is motivated in M, then either φ is in all models of TI, and so $\mathtt{TI} \supset \varphi$, or there is a causal rule of the form $\alpha \wedge \beta \supset \varphi$ in T, α is motivated in $\mathcal{M}$, and $\mathcal{M} \models \beta$. But then $\mathsf{time}(\alpha) < \mathsf{time}(\varphi)$, so we can show recursively that $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\alpha)$. Further, $\mathcal{M} \models \beta$ means that if $\varphi \in \mathtt{A}$, $\beta \in \mathtt{B}$ and $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \supset \beta$; but $\mathcal{M} \models \varphi$, φ an action, means that $\varphi \in \mathtt{A}$.

Lemma 3 *Let* $<\mathtt{A},\mathtt{B}>$ *be an occurrence kernel acceptable and total for a theory instantiation* TI. *Then if* $\mathcal{M}$ *is a model for* $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}}$, $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$ *iff* φ *is motivated in* $\mathcal{M}$.
Proof.

For every action φ, if $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \varphi$ and $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$ then φ is motivated in $\mathcal{M}$:

If $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$, then either $\mathtt{TI} \supset \varphi$, and so φ is in all models of TI, or there is a causal rule of the form $\alpha \wedge \beta \supset \varphi$ in T, $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\alpha)$, and $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \supset \beta$. But then $\mathsf{time}(\alpha) < \mathsf{time}(\varphi)$, so we can show recursively that α is motivated in $\mathcal{M}$; further, $\mathtt{TI} \supset \beta \Rightarrow \mathcal{M} \models \beta$, since $\mathcal{M} \models \mathtt{TI}$.

For every action φ, if $\mathcal{M} \models \varphi$ and φ is motivated in $\mathcal{M}$ then $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\varphi)$:

If φ is motivated in M, then either φ is in all models of TI, and so $\mathtt{TI} \supset \varphi$, or there is a causal rule of the form $\alpha \wedge \beta \supset \varphi$ in T, α is motivated in $\mathcal{M}$, and $\mathcal{M} \models \beta$. But then α contains only occurrence terms, so $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \vdash \alpha$ if $\mathcal{M}$ does, and $\mathsf{time}(\alpha) < \mathsf{time}(\varphi)$, so we can show recursively that $\mathsf{mot}(\mathtt{A},\mathtt{B},\mathtt{TI},\alpha)$. Further, since $<\mathtt{A},\mathtt{B}>$ is total for TI, $\mathcal{M} \models \beta$ means that if $\varphi \in \mathtt{A}$, $\beta \in \mathtt{B}$ and $<\mathtt{A},\mathtt{B}>_{\mathtt{TI}} \supset \beta$; but φ an action and $\mathcal{M} \models \varphi$ means that $\varphi \in \mathtt{A}$.