Motivated Action Theory:

A Formal Theory of Causal Reasoning

Lynn Andrea Stein*

Artificial Intelligence Laboratory

Massachusetts Institute of Technology

Cambridge, Massachusetts

Leora Morgenstern

IBM T. J. Watson Research Center

Yorktown Heights, New York

Abbreviated Title: Motivated Action Theory

Abstract

When we reason about change over time, *causation* provides an implicit preference: we prefer sequences of world states in which one world state leads causally to the next, rather than sequences in which one world state follows another at random and without causal connections. In this paper, we explore the crucial role that causation plays in our intuitions about temporal reasoning. We examine previous approaches to general temporal reasoning, and their shortcomings, in light of this analysis. We present a new system for *causal reasoning*, motivated action theory, which builds upon causation as a crucial preference criterion. Motivated action theory solves a broad class of temporal reasoning problems, including the traditional problems of both forward and backward reasoning, and additionally provides a basis for a new theory of explanation.

*Correspondence to 545 Technology Square Room 811, Cambridge MA 02138; (617) 253-2663; las@ai.mit.edu

1 Introduction

In this paper, we explore temporal reasoning: reasoning about how things change over time. We concentrate on temporal reasoning problems presented as descriptions of events, or "stories". Temporal ambiguity arises when multiple sequences of world states are consistent with the description presented. For example, leaving an unlocked car with the keys in the ignition leaves open (i.e. ambiguous) the question of whether it will be there in an hour, particularly in the presence of known car thieves. The possible sequences of world states here include those in which the car remains where it was left and those in which the thief absconds with it.

The space of possible interpretations, then, is the set of world-state sequences consistent with the problem description. Temporal reasoning provides an implicit preference over these sequences of world states in the form of *causation*: we prefer sequences of world states in which one world state leads causally to the next, rather than sequences in which one world state follows another at random and without causal connection. For example, we prefer sequences in which the thieves *steal* the car to sequences in which the car disappears without explanation, although not necessarily to sequences in which the car remains untouched.

By casting temporal reasoning problems in this framework, we can extract the critical insights underlying previous logical formalisms. This allows us to make direct comparisons of these principles, rather than the more common *ad hoc* comparison using a few benchmark examples. Further, by treating temporal theories uniformly, we identify both their strengths and their weaknesses; by using the first while avoiding the second, we are ultimately able to construct a new and stronger theory of temporal reasoning.

The framework that we use here treats the underlying temporal language as uniform, with temporal theories varying in the preference criteria used to select sequences of world states. We begin, in section 2, with a brief exploration of this approach to temporal reasoning. We describe

both a formal vocabulary for temporal reasoning and the temporal reasoning problems with which we shall be concerned.

In section 3, we turn to a review of early approaches to the problem. The initial discussion of temporal reasoning was in a monotonic framework. Researchers acknowledged the need for some form of *defeasible* temporal reasoning; however, the combination of temporal and nonmonotonic reasoning proved problematic. We examine Hanks and McDermott's Yale shooting problem, which describes the results of these naive extensions of nonmonotonic reasoning to reasoning over time. The Yale shooting problem demonstrates that the early approaches introduce unexpected ambiguities, in the form of implausible sequences of world states. We conclude this section with the claim that nonmonotonic temporal theories must provide some notion of *causation*.

In section 4, we use this principle—that causation provides a disambiguating preference over possible sequences of world states—to analyze several previous approaches to temporal reasoning. By focusing on their approximations of causation, we assess the extent to which these approaches adequately resolve temporal ambiguities.

The deficiencies of previous theories lead us to present *motivated action theory*, a theory of defeasible temporal reasoning which is based directly on a theory of causation. Because it relies directly on causation, motivated action theory handles the full range of temporal reasoning problems described in section 2 without reliance on limited temporal ontologies such as the situation calculus. Its causal nature further provides the basis for a theory of explanation. We initially define motivated action theory as a model-theoretic preference-based logic. After demonstrating the adequacy of the theory, we present an equivalent proof-theoretic definition together with soundness and completeness results.

2 Reasoning over Time

Temporal reasoning is reasoning about change over time. Typically, temporal reasoning problems are phrased as a set of action occurrences and state descriptions, coupled with some background knowledge about how actions cause change. The background, or domain knowledge, remains constant, while the events and states vary from scenario to scenario. For example, a *prediction* problem may involve description of an initial world state and a sequence of actions taking place in or after that initial world state. Using the background knowledge, a reasoner is expected to predict the results of performing these actions, or to describe certain details of the resulting state.

prediction

A simple prediction problem might be expressed as follows:

> A line of dominos is arranged on the table. Someone knocks over the first domino.

The domain knowledge here includes facts about one domino knocking down the next. The expected answer involves recognizing that the entire line of dominos falls down. If, for instance, the last domino's fall will cause a bell to ring, the expected outcome includes the ringing of the bell. An alternate—and less plausible—sequence of world states is one in which the line of dominos stops falling half way. While this is certainly *possible* given the initial description, it is not the sequence of world states expected to follow given that description.

In contrast, *backwards projection* problems take the form of "what is missing" queries: given some result, the reasoner is expected to identify an action or state which could have led to that action.

backwards projection

> Again, a line of dominos is arranged on the table, and someone knocks down the first.
>
> This time, the last domino does not fall. What happened?

Since this is not the expected sequence of world states, some further facts about the world must be inferred to make this sequence plausible. Depending on the background knowledge, an answer

such as "someone stopped the dominos" or "they were too far apart" might be expected. By filling in the missing information—making this sequence plausible—further projections may be made.

In both of these types of problems, the facts of the scenario together with the background knowledge delimit possible sequences of world states. In the next section, we describe a formal vocabulary for talking about temporal reasoning problems.

2.1 The Temporal Language

In this section, we describe a language for temporal scenarios. The language itself is not a logic, in the sense that it provides no inference rules and therefore has no interpretation. We give intended interpretations for some of the terms of our language, but we leave it to later sections—which describe various theories of temporal reasoning—to enforce these interpretations through particular rules of inference.

We have borrowed much of this language from Hanks and McDermott's [11] presentation of McDermott's temporal logic [24], although we have taken several liberties with that language. The ontology also incorporates a few features of McCarthy and Hayes's situation calculus [23]. However, while the situation calculus takes actions and action occurrences to be fundamental, our language—like McDermott's—is built with time points as primitive. As a result, we allow any number of actions to occur between world states. In this respect, our language resembles the one defined by Haugh in [14].

Several considerably more sophisticated temporal ontologies have been described in the literature (e.g., Allen's interval logic [4]; Hayes's histories [15]; McDermott's full temporal logic [24]; Shoham's modal system [33]). However, the naive ontology that we present here is sufficient to describe the salient features of most nonmonotonic approaches to temporal reasoning, and to demonstrate our claims with respect to the importance of causation. Indeed, the problems that

arise in this ontology would only worsen in a more sophisticated logic, and the need for some adequate notion of causation would only be strengthened.

world state

In our ontology, a point in time defines a particular *world state*. This world state is expressed as a set of state/value pairs: $\langle \mathsf{alive}, \top \rangle$; $\langle \mathsf{on(a,b)}, \bot \rangle$; $\langle \mathsf{color(house)}, \mathsf{red} \rangle$. Although the complete state of the world can be expressed by enumerating these pairs, in general we only want to describe a portion of this state. We use the notation HOLDS(t,state) to mean that state has the value $\top$ in the world state with index t. We introduce some syntactic sugar: we define $\neg\mathsf{HOLDS}(\mathsf{t},\mathsf{state}) \triangleq \mathsf{HOLDS}(\mathsf{t},\mathsf{not}(\mathsf{state})) \triangleq$ state has the value $\bot$ in the world state with index t; also $\mathsf{HOLDS}(\mathsf{t},\mathsf{state}_1 \ \& \ \mathsf{state}_2) \triangleq \mathsf{HOLDS}(\mathsf{t},\{\mathsf{state}_1,\mathsf{state}_2\}) \triangleq [\mathsf{HOLDS}(\mathsf{t},\mathsf{state}_1) \wedge \mathsf{HOLDS}(\mathsf{t},\mathsf{state}_2)] \triangleq \{\mathsf{HOLDS}(\mathsf{t},\mathsf{state}_1), \mathsf{HOLDS}(\mathsf{t},\mathsf{state}_2)\}$.[1] By a slight abuse of notation, we use "predicate" notation to express states with non-boolean value: $\mathsf{HOLDS}(\mathsf{t},\mathsf{color}(\mathsf{house},\mathsf{red}))$ means color(house) has the value red in the world state with index t, and HOLDS(t,not(color(house,red))) if the value of color(house) is not red in the world state with index t.[2] We say that $\mathsf{TIME}(\mathsf{HOLDS}(\mathsf{t},\mathsf{state})) = \mathsf{t}$; similarly, $\mathsf{STATE}(\mathsf{HOLDS}(\mathsf{t},\mathsf{state})) = \mathsf{state}$.

HOLDS

TIME

STATE

action

The state of the world is changed by *actions*. For example, if a load action occurs in a world state—with index t—in which a gun is not loaded ($\neg\mathsf{HOLDS}(\mathsf{t},\mathsf{loaded})$), then $\mathsf{t}+1$ is a world state in which the gun *is* loaded ($\mathsf{HOLDS}(\mathsf{t}+1,\mathsf{loaded})$). Although we index world states by integers, we do not insist that there be a fixed time interval between world states. For example, the time elapsed between t_0 and t_1 may not equal the time elapsed between t_1 and t_2. We use the notation OCCURS(t,act) to mean that action act occurs in the world state with index t; the resulting world state is $\mathsf{t}+1$. While actions provide transitions over world states, we do not insist that a single

OCCURS

[1] Throughout this paper, we treat sets and conjunctions interchangeably.

[2] The careful reader will note that color(house,scarlet) will yield not(color(house,red)), even if red $\triangleq$ scarlet. We can fix this by allowing multivalued, or set-valued, state; or by rigid designators; or by treating color(house,red) as a boolean-valued state (where $\mathsf{color}(\mathsf{house},\mathsf{red}) \triangleq \mathsf{color}(\mathsf{house},\mathsf{scarlet})$). In any case, these details are not important for the discussion at hand.

action occur in every world state. That is, we allow both concurrent actions—two or more actions simultaneously providing a transition between world states t and $\mathsf{t}+1$—or no action at all. When two or more actions occur concurrently, they are constrained to take the same amount of time. The result of no action in a world state is presumably a world state very much like the previous one, although time has changed, the earth has rotated, etc. We also allow statements of the form $\neg\mathsf{OCCURS}(\mathsf{t},\mathsf{act})$ that explicitly exclude any occurrence of act in the world state with index t. We define $\mathsf{TIME}(\mathsf{OCCURS}(\mathsf{t},\mathsf{act}))=\mathsf{t}$ and $\mathsf{ACT}(\mathsf{OCCURS}(\mathsf{t},\mathsf{act}))=\mathsf{act}$. ACT

To connect world states and actions, we introduce the notation CAUSES

$$\mathsf{CAUSES}(\mathsf{preconditions},\mathsf{excluded\_actions},\mathsf{actions},\mathsf{effect}) \tag{1}$$

Intuitively, this means that if preconditions hold when actions (but not excluded_actions) occur, then effect will hold (or occur) in the resulting world state. We allow preconditions to be a set (i.e., a conjunction), so that we can have multiple preconditions. Excluded_actions and actions are always sets, though excluded_actions will generally be empty and actions generally singleton. (See below.) Multiple consequences can be represented using several CAUSES statements. Effect may be either a state or an action. For example, the definition of blocks world's move might read

$$\mathsf{CAUSES}(\{\mathsf{clear(a)},\mathsf{clear(b)}\},\{\},\{\mathsf{move(a,b)}\},\mathsf{on(a,b)}) \tag{2}$$

The excluded_actions are relevant only in cases of interfering concurrent actions. For example, consider (a slight modification of) Lifschitz *et al.*'s [9] example of lifting a table that is holding a bowl of soup. Lifting only one side of the table will cause the table to tilt and the soup to be spilt. If both sides of the table are simultaneously lifted, however, this will cancel the tilting effect. We

can formalize this as:

$$\begin{array}{l}\mathsf{CAUSES}(\mathsf{on}(\mathsf{soup},\mathsf{table}),\{\mathsf{lift\_left\_side}(\mathsf{table})\},\{\mathsf{lift\_right\_side}(\mathsf{table})\},\mathsf{spilt}(\mathsf{soup}))\\ \mathsf{CAUSES}(\mathsf{on}(\mathsf{soup},\mathsf{table}),\{\mathsf{lift\_right\_side}(\mathsf{table})\},\{\mathsf{lift\_left\_side}(\mathsf{table})\},\mathsf{spilt}(\mathsf{soup}))\end{array} \tag{3}$$

but not

$$\mathsf{CAUSES}(\mathsf{on}(\mathsf{soup},\mathsf{table}),\{\},\{\mathsf{lift\_right\_side}(\mathsf{table}),\mathsf{lift\_left\_side}(\mathsf{table})\},\mathsf{spilt}(\mathsf{soup})) \tag{4}$$

Note that the problem of determining which of these actions do in fact occur is left to the temporal logic, rather than the specification language.

Since concurrent actions play little part in this paper, we will adopt a notational convenience. We introduce a three-argument version of CAUSES, which takes a singleton action (dropping the set notation for actions) and omits the excluded_actions argument entirely. Thus

$$\mathsf{CAUSES}(\mathsf{preconditions},\mathsf{act},\mathsf{state}) \tag{5}$$

should be interpreted as (and can be macro-expanded to)

$$\mathsf{CAUSES}(\mathsf{preconditions},\{\},\{\mathsf{act}\},\mathsf{state}) \tag{6}$$

We will make use of this abbreviated version of CAUSES freely and without remark in the remainder of this paper. However, our discussion below—and, in particular, Motivated Action Theory, introduced in section 5—applies equally to the case of concurrent actions.

Particular logics for temporal reasoning may enforce the use of CAUSES in different ways.

Some logics take CAUSES statements as primitive and have rules to generate inferences from these CAUSES statements. We refer to such a process as *compilation*. For example, statements of the form (5) might compile to

$$\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t}, \mathsf{preconditions}) \wedge \mathsf{OCCURS}(\mathsf{t}, \mathsf{act}) \supset \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{state})^3 \tag{7}$$

or—as we assume in section 5—

$$\begin{aligned} &\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t}, \mathsf{preconditions}) \wedge \mathsf{OCCURS}(\mathsf{t}, \mathsf{act}) \\ &\quad \wedge [\forall \mathsf{act}', \mathsf{preconditions}'. \\ &\qquad (\mathsf{CAUSES}(\mathsf{preconditions}', \mathsf{act}', \neg \mathsf{state}) \wedge \mathsf{HOLDS}(\mathsf{t}, \mathsf{preconditions}') \\ &\qquad\quad \supset \neg \mathsf{OCCURS}(\mathsf{t}, \mathsf{act}'))] \\ &\quad \supset \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{state}) \end{aligned} \tag{8}$$

Alternately, axioms such as (7) or (8) may be used to generate CAUSES statements, as Baker

[3]The compilation rules that we give in this paragraph are for the abbreviated form of CAUSES. For the full form introduced in (1), similar rules apply. For example, (7) is replaced by the following:

If effect is a state, i.e., something which HOLDS,

$$\begin{aligned} &\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t}, \mathsf{preconditions}) \\ &\quad \wedge [\forall \mathsf{ea} \in \mathsf{excluded\_acts}.\neg \mathsf{OCCURS}(\mathsf{t}, \mathsf{ea})] \\ &\quad \wedge [\forall \mathsf{a} \in \mathsf{actions}.\mathsf{OCCURS}(\mathsf{t}, \mathsf{a})] \\ &\qquad \supset \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{effect}) \end{aligned}$$

If effect is an action, i.e., something which OCCURS,

$$\begin{aligned} &\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t}, \mathsf{preconditions}) \\ &\quad \wedge [\forall \mathsf{ea} \in \mathsf{excluded\_acts}.\neg \mathsf{OCCURS}(\mathsf{t}, \mathsf{ea})] \\ &\quad \wedge [\forall \mathsf{a} \in \mathsf{actions}.\mathsf{OCCURS}(\mathsf{t}, \mathsf{a})] \\ &\qquad \supset \mathsf{OCCURS}(\mathsf{t}+1, \mathsf{effect}) \end{aligned}$$

while the form equivalent to (8) for (1) where effect is a state would be:

$$\begin{aligned} &\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t}, \mathsf{preconditions}) \\ &\quad \wedge [\forall \mathsf{ea} \in \mathsf{excluded\_actions}.\neg \mathsf{OCCURS}(\mathsf{t}, \mathsf{ea})] \\ &\quad \wedge [\forall \mathsf{a} \in \mathsf{actions}.\mathsf{OCCURS}(\mathsf{t}, \mathsf{a})] \\ &\quad \wedge [\forall \mathsf{act}', \mathsf{preconditions}', \mathsf{ex\_acts}'. \\ &\qquad [\neg \mathsf{CAUSES}(\mathsf{preconditions}', \mathsf{ex\_acts}', \mathsf{acts}', \neg \mathsf{effect}) \\ &\qquad \vee \neg \mathsf{HOLDS}(\mathsf{t}, \mathsf{preconditions}') \\ &\qquad \vee [\exists \mathsf{ea}' \in \mathsf{ex\_acts}'.\mathsf{OCCURS}(\mathsf{t}, \mathsf{ea}')] \\ &\qquad \vee [\exists \mathsf{a}' \in \mathsf{acts}'.\neg \mathsf{OCCURS}(\mathsf{t}, \mathsf{a}')]] \\ &\quad \supset \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{effect}) \end{aligned}$$

Since non-actions aren't things (i.e. can't be caused), the action case of the compilation of (1) paralleling (8) is the same as the action case of the compilation paralleling (??).

and Ginsberg [6] do. (See section 4.2, below.) In any case, each logic must somehow translate the intuitions expressed by CAUSES into axioms or rules of inference. We call these translations *causal rules*. *causal rules*

In addition to causal rules, which tell us what changes between one world state and the next, a temporal reasoning formalism must somehow enforce the *persistence* of those facts that do not change. We discuss some details of this problem, called the *frame problem*, below.[4] However, the basic issue can be stated in terms of the notation that we have already introduced: If HOLDS(t,state), how do we determine whether HOLDS(t′,state), for some t′ > t? Rules—defeasible or deductive—which enforce this constraint are called *persistence rules*. *persistence r*

2.2 Temporal Reasoning Problems

The temporal reasoning problems we have described include two parts: a particular description of world states and events, and a "background" causal theory against which this description is to be evaluated. In general, a single background theory provides the temporal model for several "stories," or scenarios; a reasoner may well have a fixed temporal theory representing its "understanding" of causality. We refer to this background information as the *theory*, and to the particular scenario as the *chronicle description*. A theory and a chronicle description together are known as a *theory instantiation*, *TI*. *TI*

The chronicle description, *CD*, is a set of specific HOLDS or OCCURS statements. Intuitively, it represents a description of some particular scenario. It may include a partial or complete description of an initial world state or of various states at later time points. It may also list some set of actions *CD*

[4] For a more extensive discussion of the frame problem, see, e.g., Brown [7] or Ford and Hayes [8].

Figure 1: A blocks-world scenario.

that occur. For example, the blocks world scenario in figure 1 is completely described by

$$\begin{array}{ll} \mathsf{HOLDS}(1,\mathsf{on}(\mathsf{a},\mathsf{b})) & \mathsf{HOLDS}(1,\mathsf{on}(\mathsf{b},\mathsf{c})) \\ \mathsf{HOLDS}(1,\mathsf{on}(\mathsf{c},\mathsf{table}_1) & \mathsf{HOLDS}(1,\mathsf{on}(\mathsf{d},\mathsf{table}_2)) \\ \mathsf{HOLDS}(1,\mathsf{clear}(\mathsf{a})) & \mathsf{HOLDS}(1,\mathsf{clear}(\mathsf{d})) \end{array} \tag{9}$$

At a later point, we may know that

$$\mathsf{HOLDS}(7,\mathsf{clear}(\mathsf{b})) \tag{10}$$

A description of events in the world state described by (9) might include

$$\mathsf{OCCURS}(1,\mathsf{move}(\mathsf{a},\mathsf{d})) \wedge \mathsf{OCCURS}(3,\mathsf{move}(\mathsf{b},\mathsf{a})) \tag{11}$$

or

$$\exists \mathsf{t} > 1.\mathsf{OCCURS}(\mathsf{t},\mathsf{move}(\mathsf{d},\mathsf{a})) \vee \mathsf{OCCURS}(\mathsf{t},\mathsf{move}(\mathsf{a},\mathsf{d})) \tag{12}$$

The sentences of CD contain no universally quantified temporal variables.[5]

The background theory, T, includes the "generic" knowledge which is true in every world state. This may include CAUSES statements, causal and persistence rules, and axioms describing other generic relationships: $\mathsf{HOLDS(t, alive) \equiv HOLDS(t, not(dead))}$, for example. A blocks world background theory might include rules such as

$$\begin{aligned}\forall \mathsf{t, a, b.HOLDS(t, clear(a)) \wedge HOLDS(t, clear(b))} \\ \wedge \mathsf{OCCURS(t, move(a, b)) \supset HOLDS(t + 1, on(a, b))}\end{aligned} \tag{13}$$

and

$$\forall \mathsf{t, b.HOLDS(t, clear(b)) \equiv [\forall a.\neg HOLDS(t, on(a, b))]} \tag{14}$$

A rule for the persistence of death might say

$$\forall \mathsf{t, a, p.HOLDS(t, dead) \supset HOLDS(t + 1, dead)} \tag{15}$$

We will say more about persistence rules below.

There are several types of temporal reasoning problems that we shall consider below. One major distinction that can be made concerns the relationship between the times about which we are given information and the time about which we must derive information. If a temporal reasoning problem requires us to describe some aspect of a world state later than any time point in CD, the problem is one of *prediction*. If the query concerns some intermediate point in CD, or some point earlier than any occurring in CD, the problem is one of *backwards projection*. Typically, backwards projection

[5] Sentences with universally quantified temporal variables are general laws which should appear in T. Examples include rules (13)–(13). Sentences which quantify over fixed intervals, such as "That March, she lived in Paris," may either be formalized as $\mathsf{\forall t, t\text{-}Mar\text{-}1\text{-}94 < t < t\text{-}Mar\text{-}31\text{-}94.HOLDS(t, Lives(Cecilia, Paris))}$ and included in T or treated as abbreviations for finite conjunctions in CD.

problems have proved difficult for temporal reasoning systems that make overly strong assumptions about the structure of events and time. We discuss some such systems in section 4.

Temporal reasoning, then, can be seen as the problem of deducing which sequences of world states "make sense." Our thesis is that we prefer sequences of world states that accord with our notion of *causation*; that is, where one world state leads to another causally, rather than those in which world states follow one another at random. In the remainder of this paper, we describe various attempts to define this preference formally.

3 Nonmonotonic Reasoning and Time

One of the earliest approaches to temporal reasoning in artificial intelligence is that of McCarthy and Hayes [23]. They introduce the *situation calculus* as a formalism for describing actions over time. As a consequence of their formalization of temporal reasoning, they discovered the *frame problem*: knowing what is true in a world state and knowing what action has taken place does not necessarily mean that we know what is true in the resulting world state. For example, the situation calculus as originally defined lacks any way to verify that moving block a onto block b does not change the location of block c. McCarthy and Hayes propose that this *frame problem* can be solved through the use of monotonic frame axioms. These axioms formalize the idea that only those states that are explicitly changed by an action change when that act is performed.

frame problem

frame problem

Unfortunately, frame axioms are not an adequate solution to the frame problem. McDermott [25, 26] has observed that, as naively implemented—e.g., asserting that c's location does not change during move(a,b)—frame axioms are simply false. For example, c might really be another name for a. Or a and c might be connected, so that moving a might force c to move as well. (This variation is due to Ginsberg and Smith [10].) Or—if concurrent actions are allowed—while we move a onto b, someone else might move c. In short, we cannot *a priori* guarantee that the location of c

will remain unchanged in the world state after move(a,b).

Essentially, frame axioms are an attempt to capture the following *law of inertia*:

$$\begin{aligned}
&\forall \mathsf{t}, \mathsf{state} \\
&\quad [\forall \mathsf{act}, \mathsf{preconditions} \\
&\quad\quad [(\neg \mathsf{CAUSES}(\mathsf{preconditions}, \mathsf{act}, \mathsf{not}(\mathsf{state}))) \\
&\quad\quad\quad \vee (\neg \mathsf{OCCURS}(\mathsf{t}, \mathsf{act})) \vee (\neg \mathsf{HOLDS}(\mathsf{t}, \mathsf{preconditions}))] \\
&\quad\quad\quad\quad \supset (\mathsf{HOLDS}(\mathsf{t}, \mathsf{state}) \supset \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{state}))]
\end{aligned} \tag{16}$$

This says that if either (1) there is no causal rule yielding not(state) or (2) the action causing not(state) doesn't occur or (3) its preconditions aren't satisfied, then state persists.

In fact, in general we want to ensure the stronger condition

$$\begin{aligned}
&\forall \mathsf{t}, \mathsf{state} \\
&\quad [\forall \mathsf{act}, \mathsf{preconditions} \\
&\quad\quad [(\neg \mathsf{CAUSES}(\mathsf{preconditions}, \mathsf{act}, \mathsf{not}(\mathsf{state}))) \\
&\quad\quad\quad \vee (\neg \mathsf{OCCURS}(\mathsf{t}, \mathsf{act})) \vee (\neg \mathsf{HOLDS}(\mathsf{t}, \mathsf{preconditions}))] \\
&\quad\quad\quad\quad \supset (\mathsf{HOLDS}(\mathsf{t}, \mathsf{state}) \equiv \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{state}))\ ]
\end{aligned} \tag{17}$$

The contraposed consequence—$\mathsf{HOLDS}(\mathsf{t}+1, \mathsf{state}) \supset \mathsf{HOLDS}(\mathsf{t}, \mathsf{state})$, given the same antecedent—is the *law of causation*, and ensuring it is the point of causal theories. To make the law of causation hold (given unexpected results), we can either postulate new CAUSES statements—as do Lifschitz, Baker and Ginsberg, etc. (see section 4.2)—or derive new preconditions—Lifschitz does this, too—or assume that additional actions must have occurred. This last is the approach that we adopt in motivated action theory (section 5).[6]

[6]Note, however, that the empty act is possible, allowing spontaneous actions (see section 5.3). Motivated action

In any case, simply adding an axiom like inertia—(16)—provides no better solution than frame axioms: we still need to *defeasibly* rule out unexpected causal connections (clause (1)) and action occurrences (clause (2)), and we need to make default assumptions about state (clause (3)), particularly when we have incomplete information. Nonetheless, inertia—and frame axioms—point in the right direction, and we will return to them below.

These early attempts to formalize temporal reasoning led to the realization that some form of nonmonotonic reasoning would be necessary. For example, McDermott assumes some appropriate nonmonotonic logic in describing his temporal logic [24]; McCarthy presents temporal reasoning as an application of circumscription [21, 22]; Reiter uses temporal reasoning as a motivating example for his default logic [30]. Indeed, early approaches to both nonmonotonic and temporal reasoning simply assumed that it would eventually be possible to take a suitable temporal logic and "plug in" some nonmonotonic logic to achieve nonmonotonic temporal reasoning.

In [11, 12, 13], Hanks and McDermott present the Yale shooting problem. This seemingly simple temporal reasoning problem proves notoriously difficult for classical nonmonotonic logics. The Yale shooting problem—restated in our temporal vocabulary—consists of the chronicle

$$\mathsf{HOLDS}(1, \mathsf{alive}) \wedge \mathsf{HOLDS}(1, \mathsf{loaded}) \wedge \mathsf{OCCURS}(2, \mathsf{shoot}) \tag{18}$$

coupled with the background theory

$$\mathsf{CAUSES}(\mathsf{loaded}, \mathsf{shoot}, \mathsf{not}(\mathsf{alive})) \tag{19}$$

theory applies the equivalent of a closed world assumption to assume that actions occur spontaneously only if they are explicitly so described.

which—according to their definition of causation—compiles as described on page 9 into

$$\forall t.\mathsf{OCCURS}(t, \mathsf{shoot}) \wedge \mathsf{HOLDS}(t, \mathsf{loaded}) \supset \neg\mathsf{HOLDS}(t+1, \mathsf{alive}) \tag{20}$$

and persistence rules indicating that $\mathsf{loaded}, \mathsf{alive}, \neg\mathsf{loaded}$, and $\neg\mathsf{alive}$ continue to hold (unless explicitly changed). We deliberately omit the particular form of these persistences, since they depend on the nonmonotonic logic in which the Yale shooting problem is expressed. Hanks and McDermott give such persistence rules and demonstrate that this anomaly arises for three "standard" nonmonotonic logics: McCarthy's circumscription [21], McDermott and Doyle's nonmonotonic logic [27], and Reiter's Default Logic [30].

We would like to predict $\neg\mathsf{HOLDS}(3, \mathsf{alive})$. Relative to the standard non-monotonic logics, however, the chronicle description supports (at least) two sequences of world states: the expected one, in which one reasons by default that $\mathsf{HOLDS}(2,\mathsf{loaded})$, and in which $\neg\mathsf{HOLDS}(3, \mathsf{alive})$; and an unexpected sequence in which one reasons from the persistence rules that $\mathsf{HOLDS}(3,\mathsf{alive})$ and in which, therefore, $\neg\mathsf{HOLDS}(2, \mathsf{loaded})$. Standard non-monotonic logic gives us no way of preferring the expected—intuitively correct—sequence of world states to the unexpected one.

Hanks and McDermott argue that the Yale shooting problem sounds the death knell for nonmonotonic logics. They claim that the inability of general-purpose nonmonotonic logics to resolve basic temporal ambiguities proves that nonmonotonic logic as an endeavor is doomed to failure. Fortunately, the perspective of time allows us to better understand and restate their pessimistic conclusions.[7] The issue in the Yale shooting problem is not that general nonmonotonic logics can't do general nonmonotonic reasoning, but that temporal reasoning is *not* general nonmonotonic reasoning. Temporal reasoning involves a specific kind of ambiguity—temporal ambiguity, or

[7] To be perfectly fair, even Hanks and McDermott [12, 13] don't agree with Hanks and McDermott [11]; in their later writings, they agree that their initial conclusions were overly pessimistic.

ambiguity over sequences of world states—and temporal ambiguity comes tailor-made with its own preference criterion: causation. When reasoning about temporal problems, we prefer sequences of world states in which one world state leads causally to the next—as in the expected model—rather than sequences in which one world state follows another at random and without causal connection—as when the gun becomes, inexplicably, unloaded. The problem with using general-purpose nonmonotonic logics to perform temporal reasoning is that these logics contain no inherent notion of causation. In [34], we look at several forms of reasoning and the ways in which they differ from general nonmonotonic reasoning. Within each form of reasoning, a specific preference criterion resolves ambiguity. The preference criterion is an essential part of the form of reasoning: causality in temporal reasoning, specificity in taxonomic (inheritance) reasoning, and similarity in counterfactual reasoning. Various formal theories within each of these forms of reasoning are analyzed by restating them in terms of a single language (as with the temporal language of section 2.1) and comparing their preference criteria to one another and to the ideal inherent in the form of reasoning.

In the next section, we look at several attempts to solve the Yale shooting problem. Each works, in some sense, by trying to build a notion approximating causation into nonmonotonic logic. By comparing their approximations of causation with our intuitions, we can see the extent to which these solutions succeed and the extent to which they fall short of our expectations. In section 5, we present motivated action theory, a theory of causal reasoning whose development was motivated by this view that causation provides an ambiguity-resolving preference over sequences of world states.

4 Comparisons of Existing Theories

In this section, we examine several previous theories of temporal reasoning. In the past, these comparisons have been made by reference to a series of "benchmark problems." Each new paper

posed a benchmark problem that no previous logic can solve and presents a logic to solve it. Rather than add to the existing collection of benchmark problems, we will compare the principles underlying the logics themselves. We do this by interpreting the logics in a single framework—the temporal language of section 2.1—and separating out their inference processes—or preference criteria—from the underlying language. As a result, we can directly examine their model selection criteria and compare them to the fundamental preference of temporal reasoning: causation. By defining them in terms of their underlying preferences over the space of possible sequences of world states, we make possible a principled comparison between disparate techniques. Similarly, by comparing these preferences to our intuitive notion of causation, we can see where and why these previous approaches succeed as well as identify their inevitable weaknesses.

These first attempts to solve the Yale shooting problem are often divided into two categories: those that concentrate on the structure of time, and those that focus on cause-and-effect. The two approaches appear divergent; nonetheless, motivated action theory (section 5) can be seen as a successor to both.

4.1 Chronological Approaches

Hanks and McDermott [11, 12, 13] argue that the problem with general nonmonotonic logics is their failure to incorporate the notion of time. In particular, they claim that time creates an explicit ordering, and temporal reasoning is inherently biased towards that ordering. In the Yale shooting problem, the expected model arises when we reason about world states in temporal order: alive and loaded hold at 1, so they will (by default) hold at 2. This means that when the gun is fired (at 2), loaded holds, clipping alive (so $\neg$HOLDS(3, alive)). In contrast, the unexpected model arises when we apply persistence to alive—yielding HOLDS(2,alive) and HOLDS(3,alive)—before we have reached any conclusion about loaded at 2. HOLDS(3,alive) and OCCURS(2,shoot) now force us to

reason *backwards* about loaded—it *must* be the case (by rule (20)) that $\neg$HOLDS(2, loaded).

Chronological solutions address this particular point. Each of these solutions—Hanks and McDermott's program [11], Shoham's logic of chronological ignorance [31, 32], Kautz's logic of persistence [16], and temporal applications of Lifschitz's pointwise circumscription [17]—describes a reasoning system with an inherent forward temporal bias. Each works by considering world states in their chronological order, extending as many persistences as possible through earlier world states before addressing later world states. This approach yields a particular preference over sequences of world states: we prefer sequences in which changes take place—persistences are clipped—in later world states, rather than earlier ones. This in turn leads to the "motto" of chronological solutions: we prefer that *as little happens for as long as possible*.

Hanks and McDermott's program works by updating world states in temporal order. Thus, in the shooting problem, it analyzes the world state at 2—by default, alive and loaded persist—and *then* the world state at 3—since OCCURS(2,shoot) and HOLDS(2,loaded), $\neg$HOLDS(3, alive). The underlying idea is to postpone changes until they are forced; or, to allow persistences to continue for as long as possible. This avoids the anomalous model which arises for the standard nonmonotonic logics.

The three chronological logical approaches essentially mimic the behavior of Hanks and McDermott's program. Kautz and Lifschitz use circumscription to fix state values in one world state before considering the next; Shoham defines a model preference criterion with the same properties. This approach prefers sequences of world states that minimize changes to the world; persistences apply whenever possible. Changes occur only when actions (with suitable preconditions) force them to happen.

Problems with Chronological Solutions For several reasons, forward reasoning solutions are not entirely satisfactory. The most obvious is that causation is not merely time-moving-forward.

For example, when we are performing the "what went wrong" type of reasoning typical of backwards projection, we reason from the appearance of an effect *backwards* in time to its possible causes. Consider, for example, a modification of the Yale shooting problem, where *CD* contains

$$\begin{array}{ll} \mathsf{HOLDS}(1, \mathsf{alive}) & \mathsf{HOLDS}(1, \mathsf{loaded}) \\ \mathsf{OCCURS}(5, \mathsf{shoot}) & \mathsf{HOLDS}(6, \mathsf{alive}) \end{array} \tag{21}$$

(we have moved the shoot to 5, and added the (unexpected) outcome that shooting did *not* lead to not(alive)). Since alive holds at 6, we know that the gun must somehow have become unloaded between times 2 and 5; however, we cannot say exactly when this happened. In contrast to this intuition, the systems of Shoham, Kautz, and Lifschitz prefer that the sequence of world states in which the gun became unloaded between time 4 and time 5. This is because that sequence postpones the change for as long as possible. Kautz first noted this point when he presented his solution to the Yale shooting problem.

This leads to a second objection to chronological solutions: they do not seem to address the real concerns underlying the Yale shooting problem. We don't reason that HOLDS(3,not(alive)) *because* we reason forward in time. We reach this conclusion because we are told of an action that causes not(alive), but are not told of any action that causes not(loaded). Chronological solutions substitute time-moving-forward for causation; but causation, not chronological reasoning, is at the heart of temporal reasoning. Chronological approaches work when their preference—changes happen later—coincides with causation. These scenarios include the original Yale shooting problem as well a larger class—described by Shoham [32]—of temporal projection problems. But where the two criteria diverge—for example, in backwards projection—chronological minimization does not provide an adequate preference criterion for resolving temporal ambiguity.

From chronological solutions to minimizing actions: moving towards MAT Here, we pause to examine the reason that chronological solutions do work for temporal projection problems. Chronological solutions minimize what is true at earlier time points, forcing truths at later points. In fact, it turns out that the truths that are minimized are only the changes—the action occurrences—and not the states. For example, in the Yale shooting problem, it's not not(alive) or not(loaded) that needs to be put off; it is the unload *action*. That is, unless we *know* that an unload occurs, we allow loaded to persist; later, when we get to the shoot, we are forced to give up the persistence of alive. We never minimize state, after all. So we can achieve the same result by preferring world states in which actions occur later. In this case, we would not add the load action, so shoot would clip alive (We can't ever *exclude* the shoot, since it is in our axiomatization. We can only block loaded, so that the shoot does not affect alive). Kautz's *logic of persistence* does exactly this, by explicitly minimizing (circumscribing) *clipping*s, or endpoints of persistences.

Actually, once we have started minimizing actions, it turns out that we don't need to minimize them chronologically at all. Minimizing actions solves a whole class of temporal reasoning problems, although two specific problems with this approach, *causal chains* and *spontaneous actions*, remain. Nonetheless, the underlying intuition forms the basis of motivated action theory (see section 5). We do not prefer that fewer actions happen *earlier*; instead, we prefer that fewer extraneous actions happen. (An extraneous action is one that is not forced, either by being explicitly mentioned in the axiomatization or by following directly from the causes mentioned in it). Circumscribing actions altogether—preferring those sequences of world states in which *as little happens*, period—solves the Yale shooting problem. Motivated action theory solves the additional problems of spontaneous actions and causal chains by excluding motivated—non-extraneous—actions from the minimization, allowing for the occurrence of *motivated* actions.

This is more in accordance with our intuitions about causation: uncaused actions do not hap-

pen. When no unmentioned actions are caused, ruling out uncaused actions reduces to ruling out unmentioned actions. When reasoning forward—*predicting*—this in turn reduces to postponing action commitments. With this analysis, we an easily see that chronological solutions will be adequate for prediction, and minimal-action models for scenarios with no unmentioned caused actions. Motivated action theory (section 5) relies on our original intuition and so will handle a still broader class of scenarios.

4.2 Causal Approaches

The situation calculus as originally conceived by McCarthy and Hayes models action occurrence as a function, which they call Result, mapping an (action, world state) pair onto a unique world state. All of the examples in the early papers on the situation calculus [20, 23] describe universes in which exactly one action happens at any time. Although the absence of concurrency is not explicitly included in the original formalization of the situation calculus, it has been implicitly or explicitly assumed in virtually all later work that concurrency is not allowed.[8] Further, transitions from one world state to the next take place only when an action occurs. In our temporal language, these restrictions may be expressed as

$$\forall \mathsf{t}, \mathsf{act}.\ \mathsf{OCCURS}(\mathsf{t}, \mathsf{act}) \quad \equiv \quad [\forall \mathsf{act}' \neq \mathsf{act}.\ \neg \mathsf{OCCURS}(\mathsf{t}, \mathsf{act}')] \tag{22}$$

Given this *situation calculus axiom*, the problem of determining that no action occurred to unload the gun becomes trivial; all actions are defined by the transitions from t to $\mathsf{t}+1$ to $\mathsf{t}+2$, etc. However, the ontology exposes a second temporal reasoning issue: the problem of determining that

[8] In fact, McCarthy and Hayes originally assumed that the situation calculus would be able to handle more complex scenarios, including concurrent actions [personal communication]. For versions of the situation calculus which explicitly permit concurrent actions see, e.g., Haugh [14], Gelfond, Lifschitz, and Rabinov [9], or the work-in-progress of McCarthy (personal communication).

the *known* actions don't have unusual effects. In the Yale shooting problem, the first of these issues involves determining that no action occurs in the world state with index 2; the second is the problem of determining that the null action—traditionally called wait—that does occur has no side effects. The solutions described in this section—Lifschitz's formal theories of action [18], Baker and Ginsberg's abnormal-for-state [6], and Haugh's causal minimizations [14]—address the second of these problems; Haugh's also addresses the first.

These solutions are not based on forward reasoning strategies. Rather, they work by circumscribing over CAUSES(..., act, state). Formally, these theories divide our predicate, CAUSES, into two predicates: precond(preconditions, act), and causes(act, state). Circumscribing causes means that we prefer sequences of world states in which actions have fewer effects. This is certainly a part of our notion of causation: actions cause only the expected—explicitly stated[9]—changes. For example, in the Yale shooting problem, circumscribing causes limits its extent to causes(shoot,not(alive)). However, this is not a complete notion of causation. For example—as we shall see—by itself this is insufficient to prevent *other* actions from (occurring and) causing changes that we don't expect.

Formal Theories of Action In Lifschitz's formal theories of action—which makes use of the original situation calculus—exactly one action occurs in each world state and that action is known. To capture the null action which occurs in the world state with index 1, a wait action is defined with no (explicit) causal consequences. Circumscribing causes now yields no *implicit* causal consequences for wait, so Lifschitz determines that nothing changes during the wait action. Thus, loaded persists, HOLDS(2,loaded), and ¬HOLDS(3, alive).

This solution doesn't force reasoning to go forward in time. Nevertheless, it is highly problematic. It depends on the situation calculus constraint, which requires the problem description to provide all and exactly those actions that do occur. Consider what would happen in a world

[9] or otherwise logically necessary

in which concurrent (or uncertain) actions were allowed and in which we were to add the rule causes(unload, not(loaded)) to the theory. We could then have a sequence of world states where OCCURS(1,unload), yielding HOLDS(3,alive). There would be no way to prefer the expected sequence, where ¬HOLDS(3, alive). This cannot in fact happen in Lifschitz's formulation because in his rigid situation calculus framework concurrent actions are not allowed. Since OCCURS(1,wait), nothing else can happen and unload actions are ruled out.

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 and prefers sequences of world states in which the given actions have only their minimally required effects. The intuition underlying the Yale shooting problem, however, is that we can make reasonable temporal projections in worlds where concurrent or unexpected actions are allowed. 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.

A second problem with this framework involves the backward reasoning scenario of the previous section. If—as in that scenario—HOLDS(6,alive), then circumscribing causes yields causes(wait, not(loaded)). That is, the null action—waiting—*causes* the gun to become unloaded. While the semantics of this statement may be unsettling, its effects are worse. Now, every time a loaded gun is left to wait, it will become unloaded: waiting *causes* unloadedness. This problem arises because causes deals with action types—shootings, unloadings, *etc.*—rather than with particular instances—e.g., the shoot in world state 5.

Fixing Formal Theories of Action Since this second objection was first noted, Lifschitz and Rabinov have constructed a theory of "miracles" to deal with it [19]. The idea here is that if we must postulate additional causes—such as the magically unloaded gun of the previous example—we can do so by allowing that a *miracle* happened rather than by assuming that the wait action *caused* the

unloading. Formally, they circumscribe both miracles and causes, but miracles are circumscribed at a lower priority than causes (so that we are more willing to admit miracles than new causes).

The miracle mechanism is actually an elaborate attempt to compensate for the inability of the situation calculus to express concurrent actions. Intuitively, the unloading that must occur (so that shooting does not cause not(alive)) is the result of some unload action. Lifschitz's original version made the unloading the result of the wait action, and further waits could therefore be expected to have the same result. Lifschitz and Rabinov make it the result of a miracle, so that it is unlikely to recur during further waits. But miracles are still not unloads.

Consider, for example, the blocks-world scenario in figure 1. If we assert that HOLDS(4,on(b,d)) without giving an explicit sequence of actions, Lifschitz would presumably formalize this as wait occurring at 1, 2, and 3. Now, certainly some "miracle" must occur to put b on d: the miracle that is equivalent to move(b,d). But in order for the move to take place, clear(b) must hold. Thus, we actually know that a has been moved.[10] In contrast, Lifschitz and Rabinov are able to assert only that by some miracle, b has come to be on top of d. This in itself may not be alarming, but now imagine that a is actually A*, the world-famous and fabulously precious diamond. Because it is so valuable, A* is attached to all of the finest alarms that money can buy. In fact, then, if HOLDS(4,on(b,d)) we can reasonably expect that all of these alarms have been tripped; Lifschitz and Rabinov can only state that a miracle occurred. Indeed, moving a without effect is nothing short of a miracle.

The fundamental problems with Lifschitz's solutions stem from the fact that it minimizes change *types*, without minimizing change *tokens*. That is, circumscribing causes minimizes the changes that a particular *kind* of action can cause, but it does not address either the changes that a particular act—an *instance* of an action—can cause, or which action instances can occur. Lifschitz's solution

[10] Inferring all of the effects of an action has been called the ramification problem.

might be paraphrased *action types cause as few changes as possible*. This solution is both important and necessary, but it is not itself sufficient.

Abnormal-for-State Baker and Ginsberg [6] suggest a solution within this single action paradigm of the situation calculus as well. However, while Lifschitz minimizes causes, Baker and Ginsberg minimize something much closer to CAUSES. That is, where Lifschitz treats causation as a property of action types, Baker and Ginsberg supply a notion of causation with *state* as an argument. They call this predicate ab_v (abnormal-for-state). Unlike the preconditions argument to CAUSES, however, the first argument to Baker and Ginsberg's $\mathsf{ab}_v(\mathsf{preconditions}, \mathsf{act}, \mathsf{state})$ must be *complete*:

$$\begin{aligned} &\forall \mathsf{preconditions}, \mathsf{act}, \mathsf{state}, \mathsf{state}'. \\ &\quad \mathsf{ab}_v(\mathsf{preconditions}, \mathsf{act}, \mathsf{state}) \\ &\qquad \supset [(\mathsf{state}' \in \mathsf{preconditions})] \vee [(\mathsf{not}(\mathsf{state}') \in \mathsf{preconditions})] \end{aligned} \tag{23}$$

This means that ab_v depends on the value of every state in the world state when act occurs.[11]

Perhaps more importantly, Baker and Ginsberg do not assume that CAUSES is a primitive notion. Instead, they derive it from descriptions of world states and the actions that connect

[11] A second difference between ab_v and CAUSES is in the third argument: while CAUSES indicates that state is to be true in the resulting world state, ab_v does not specify whether state or not(state) holds in the resulting world state; it says only that the truth value has changed from its value in preconditions. Since its value is given explicitly in ab_v's first argument, this difference is relatively insignificant.

them.[12] They assert

$$\begin{array}{l}\forall \mathsf{t}, \mathsf{act}, \mathsf{state}, \mathsf{preconditions}. \\ \quad \neg(\mathsf{ab}_v(\mathsf{preconditions}, \mathsf{act}, \mathsf{state}) \\ \quad \wedge \mathsf{OCCURS}(\mathsf{t}, \mathsf{act}) \wedge \mathsf{HOLDS}(\mathsf{t}, \mathsf{preconditions})) \\ \quad\quad \supset [\mathsf{HOLDS}(\mathsf{t}, \mathsf{state}) \equiv \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{state})]^{13}\end{array} \tag{24}$$

This is essentially a reformulation of inertia (16) and causation in the context of the situation calculus axiom (22) and completeness condition (23). For example, from the causal rule in the Yale shooting problem (20), the situation calculus axiom (22), and Baker and Ginsberg's axiom (24), we can derive $\mathsf{ab}_v(\{\mathsf{alive}, \mathsf{loaded}\}, \mathsf{shoot}, \mathsf{not}(\mathsf{alive}))$.[14]

Baker and Ginsberg's solution, then, is to minimize ab_v, their version of CAUSES. The result, as for Lifschitz's causes, is a theory that prefers those sequences of world states in which actions cause only the expected changes. Their theory does not treat concurrent actions, and so suffers from the same possible concurrent unload problem as Lifschitz's. It does behave differently with respect to wait's *causing* unloading: now wait only causes unloading in a particular state.

Causal Minimizations Haugh [14] avoids these pitfalls by explicitly allowing concurrent actions. This reopens the first Yale shooting problem: deducing that nothing else happens to unload the gun. Haugh solves both the "nothing else happens" problem and the "no bizarre side effects"

[12] The fact that Baker and Ginsberg *derive* their "causes" predicate leads Ginsberg to argue that ab_v is *not* causation [personal communication]. Indeed, Ginsberg (and presumably Baker) would object to their theory's inclusion in this section. While this point may ultimately prove to be of philosophical import, we will continue to treat ab_v as a causal predicate on the looks, walks, quacks like a duck principle.

[13] We have taken the liberty of reformulating Baker and Ginsberg's axiom in our notation; the original notation makes use of the situation calculus function result and their function describes. In the context of the situation calculus rule (22) and the completeness condition (23) our reformulation is equivalent to their (4) and (11) [6, pp. 908 and 909].

[14] Actually, we derive $\mathsf{ab}_v(\{\mathsf{alive}, \mathsf{loaded}, \mathsf{X}\}, \mathsf{shoot}, \mathsf{not}(\mathsf{alive}))$, where X represents the rest of the states in any world state: blueSky, not(blueSky), on(b,c), *etc.* This is because Baker and Ginsberg insist that the first argument to ab_v be complete.

problem by minimizing potential-causes, the conjunction of OCCURS and causes.[15] This means that (1) an action that must occur has as few effects as possible, and (2) anything with known effects occurs only if it must: prefer sequences of world states in which *the fewest changes actually happen*. Since unload is known to cause not(loaded), *if* it were to occur, it would be a potential-cause. We can't minimize causes(unload, not(loaded))—it is in our axiomatization. So instead we minimize OCCURS(...,unload), and the unload never happens.

This idea seems to merge our comments regarding the utility of minimizing actions—from section 4.1—with Lifschitz's suggestions regarding unexpected effects. Indeed, it is quite effective in many scenarios, and many of these intuitions are reflected in motivated action theory, below. Haugh's theory is unable to handle causal chains and spontaneous actions, which we describe in section 5.3, below. Here, we mention some strange results which Haugh obtains when reasoning about disjunctions.

Suppose, for example, that 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 have 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" [24, 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

[15] This causes is Lifschitz's causes(act, state), again. Haugh, too, uses a precond predicate for the other half of CAUSES. He actually presents two solutions: *potential causes*, described here, and *determined causes*. Haugh's theory of determined causes adds a chronological aspect, crossing his theory of potential causation with the chronological solutions of the previous section. The resulting theory suffers from anomalies results similar to but more severe than those described here for potential causes.

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,blow-up(world)).

5 Motivated Action Theory

We have analyzed several theories of temporal reasoning by examining the ways in which each had a flawed notion of what *causation* means as a preference over sequences of situations. This view gives a means of uniform comparison for theories originally defined in terms of diverse logics and temporal models. We have seen that chronological approaches minimize changes to the world in temporal order, allowing states to persist for as long as possible. This approach fails when causal rules are used to reason backwards in time. In contrast, causal approaches minimize the kinds of change that may occur. When coupled with the situation calculus, which insists that exactly one action occur in every situation, these solutions enjoy moderate success. However, care must be taken when combining causal solutions with a richer temporal ontology.

In this section, we describe *motivated action theory*, a theory of causal reasoning which combines features from both of these approaches but works within a richer temporal ontology and handles a broader class of problems than the previous theories. From the chronological approaches, and from

Haugh's causal minimization, we adapt the idea of minimizing actions; from the causal approaches, we borrow the idea of minimizing causes. The resulting theory depends strongly on the explicit form of the problem statement. It solves both projection and backwards reasoning problems in the context of concurrent actions, and additionally lends itself 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 rely on the idea of a *motivated action*, an action that *must* occur in a particular world model.

5.1 The Form of the Rules

For the most part, motivated action theory makes use of a language like that of section 2.1. However, motivated action theory is an explicitly syntactic theory; as a result, the *form* of causal rules plays a critical role. We discuss it briefly here.

In motivated action theory, a causal rule is a sentence of the form

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

where:

α is a set of occurrence terms $\mathsf{OCCURS}(\mathsf{t}_i, \mathsf{act}_i)$—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,

γ describes the results of the action, and

$\exists \mathsf{t}. \forall x \in \alpha\beta, \mathsf{TIME}(x) \leq \mathsf{t}$ and $\mathsf{TIME}(\gamma) \geq \mathsf{t}+1$.

Note that γ can include occurrence terms. We can thus express causal chains of actions.

This notation is not exclusive of the CAUSES notation we have described above. In fact, causal rules can be derived from CAUSES statements. For each CAUSES statement (5), add the axiom

$$\begin{aligned} &\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t}, \mathsf{precondition}) \wedge \mathsf{OCCURS}(\mathsf{t}, \mathsf{act}) \\ &\supset \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{state}) \end{aligned} \tag{25}$$

In addition, since motivated action theory allows causal chains, whenever there is a statement of the form

$$\mathsf{CAUSES}(\mathsf{preconditions}, \mathsf{act}, \mathsf{act}') \tag{26}$$

we add the axiom

$$\begin{aligned} &\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t}, \mathsf{precondition}) \wedge \mathsf{OCCURS}(\mathsf{t}, \mathsf{act}) \\ &\supset \mathsf{OCCURS}(\mathsf{t}+1, \mathsf{act}')] \end{aligned} \tag{27}$$

Conversely, we can use Baker and Ginsberg's method of deriving CAUSES from our causal rules and inertia (16), provided that we use the methods outlined below to rule out actions that don't occur. Baker and Ginsberg don't encounter this difficulty because they rely on the situation calculus to eliminate all but a single action.

We should also include a brief word on persistence rules and the qualification problem. Motivated action theory simply takes the axiom of inertia (16) as stated above. We solve the problem of proving non-occurrence of actions through motivation and preferred models; we solve the problem of unknown state by allowing state to vary freely—in one model, state may hold, while in another, not(state) does.

This leaves only the difficulty of ruling out unknown CAUSES. Here, we turn to the causal

approaches to temporal reasoning. Like Lifschitz and Baker and Ginsberg, we exploit inertia and the closed-world assumption, allowing us to compile persistence rules from causes. Before using motivation to determine preferred models as described below, we circumscribe CAUSES. Now, inertia (16) allows us to derive monotonic persistence rules. For example, a formulation of the Yale shooting problem might include the rule

$$\begin{aligned} &\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t}, \mathsf{alive}) \\ &\quad \wedge \neg\, (\mathsf{OCCURS}(\mathsf{t}, \mathsf{shoot}) \wedge \mathsf{HOLDS}(\mathsf{t}, \mathsf{loaded})) \\ &\quad \supset \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{alive}) \end{aligned} \tag{28}$$

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:[16] 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 may allow causal rules to fire as well.

5.2 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 HOLDS(t,p) or OCCURS(t,a) follows from the

[16] In section 5.4, we give a proof theoretic version of motivated action theory, but the axioms remain monotonic. There, 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. In formal terms, given $TI = T \cup CD$, we are interested in determining the preferred models for TI. $\mathcal{M}(TI)$ denotes a model for TI: i.e., $\mathcal{M}(TI) \models TI$. We define a preference criterion for models in terms of *motivated* actions: those actions which "*have to happen.*"[17] Our strategy will be to minimize those actions which are *not* motivated. (We actually define motivation over all statement types, but in the end it will only be the motivation of occurrence terms about which we care.)

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 $\mathcal{M}(TI)$ or semi-motivated in $\mathcal{M}(TI)$ or existentially motivated in $\mathcal{M}(TI)$. *motivated*

We now proceed to define the various types of motivation. To begin with, it is clear that actions that follow directly from the theory instantiation TI will "have to be" in $\mathcal{M}(TI)$, for any model $\mathcal{M}(TI)$. For example, in the original Yale shooting problem—(18) together with (20)—OCCURS(2,shoot) is in CD, so it will certainly be in $\mathcal{M}(TI) = \mathcal{M}(T \cup CD)$. This is motivation in its strongest sense.

Definition: Given a theory instantiation $TI = T \cup CD$, we say that a statement φ is *strongly motivated* with respect to TI if it is in all models of TI, i.e. if $TI \models \varphi$. *strong*

If φ is strongly motivated with respect to TI, we say that it is motivated in $\mathcal{M}(TI)$, for all models $\mathcal{M}(TI)$. *motivation*

Strong motivation includes actions that are deductive consequences of other actions (or states) as well as those actions explicitly mentioned. For example, if opening a safe inevitably causes air

[17] Amsterdam [5] has objected to our use of the phrase "has to be in that model." He points out that a statement that holds in a model trivially must be in that model, and objects that our language is therefore meaningless. We can easily brush aside his objection by protesting that we really mean "partial truth assignment," or "set of models," rather than models. But it seems to us that our use of the phrase was not careless; instead our language captures the intuition that motivated actions are somehow a more necessary part of the model, forced by the presence of their causes, than chance facts such as the color of a shirt someone happens to be wearing.

to rush in, air rushing in is strongly motivated even if it is unmentioned in a *CD* containing the safe's opening.

A weaker form of motivation occurs when an action may or may not happen. For example, a batch of freshly baked cookies in the kitchen may well be devoured by roaming cookie thieves. Our *CD* might contain $\mathsf{HOLDS}(1, \mathsf{in}(\mathsf{cookies}, \mathsf{kitchen}))$—but no information as to the presence or absence of cookie thieves—and our T might include rules such as

$$\begin{aligned} &\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t}, \mathsf{in}(\mathsf{cookies}, \mathsf{kitchen})) \\ &\wedge \mathsf{HOLDS}(\mathsf{t}, \mathsf{in}(\mathsf{cookie\text{-}thief}, \mathsf{kitchen})) \\ &\supset \mathsf{OCCURS}(\mathsf{t}+1, \mathsf{cookie\text{-}theft}) \end{aligned} \tag{29}$$

In this case, some models of TI will entail $\mathsf{HOLDS}(1, \mathsf{in}(\mathsf{cookie\text{-}thief}, \mathsf{kitchen}))$ and therefore $\mathsf{OCCURS}(2, \mathsf{cookie\text{-}theft})$, while others will entail neither. $\mathsf{OCCURS}(2, \mathsf{cookie\text{-}theft})$ is not entailed by all models of TI, and so it is not strongly motivated (in TI). It is, however, weakly motivated in $\mathcal{M}(TI)$ whenever $\mathcal{M}(TI) \models \mathsf{HOLDS}(1, \mathsf{in}(\mathsf{cookie\text{-}thief}, \mathsf{kitchen}))$. If there's a cookie thief around, the theft "has to happen."

Definition: A statement φ is *weakly motivated* in $\mathcal{M}(TI)$ if there exists in TI a causal rule as defined above; α is motivated in $\mathcal{M}(TI)$; and $\mathcal{M}(TI) \models \beta$. *weak motivation*

The α clause is added to ensure that causal consequences of motivated actions—like the falling of successive dominos—are motivated. Note that—as defined above—α may be strongly or weakly or semi- or existentially motivated.

The definition of weakly motivated depends on the form of the causal rule $\alpha \wedge \beta \supset \varphi$. Logically, this is equivalent to $\neg\alpha \vee \neg\beta \vee \varphi$, or $\alpha \wedge \neg\varphi \supset \neg\beta$, or any number of other variations. For example,

the causal rule for shoot (20) might be rewritten as

$$\forall \mathsf{t}.\mathsf{OCCURS}(\mathsf{t}, \mathsf{shoot}) \wedge \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{alive}) \supset \mathsf{HOLDS}(\mathsf{t}, \mathsf{not}(\mathsf{loaded})) \tag{30}$$

This rule says that if someone is shot, but remains alive, the gun must not have been loaded. Indeed, this statement is reasonable, and our causal rule (20) would allow such an inference (see section 5.3, below, for details). But we do *not* want to say that shooting someone who remains alive *causes* the gun to have been unloaded, so we do not include axiom (30) in our causal rules, and we do not allow it to participate in motivating other actions. We discuss this point further in [34], especially in chapter 5.

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 world state—our defeasible conclusions.

In addition to these two types of motivation, we need to define special mechanisms to handle complex expressions. Conjunctions in TI can simply be broken into independent assertions, as the entire TI is implicitly conjoined. Universal quantification is treated as follows: We say that a statement of the form $\forall x \varphi(x)$ is motivated iff for all constants c, $\varphi(c)$ is motivated. However, disjunction[18] and existential quantification are more complex, and additional machinery is required to handle these constructs.

Disjunction is treated similarly to weak motivation. Consider a baby with a plate of food in front of him. Babies being babies, he will either toss his food on the floor or make a tremendous mess of his face. In this case, food-throwing will be true in some models, while a messy face will be

[18] We are indebted to Matt Ginsberg for pointing out the difficulties that arise with disjunctions in CD.

true in others. (In some, indeed, both food-throwing and a messy face will be true.) Because we have adequate explanation for any of these consequences, we say that the food-throwing (similarly the face-messing) is motivated in any model in which it is entailed. In general, when CD contains a disjunction or when a causal rule implies a disjunction, each disjunct is motivated in the model(s) that entail(s) it.

semi-motivat

Definition: A statement φ is *semi-motivated in* $\mathcal{M}(TI)$ if it is contained in a disjunction $\rho = \psi_1 \vee \varphi \vee \psi_2$,[19] $\rho \in CD$ or there is a causal rule $\alpha \wedge \beta \supset \rho \in T$ with α motivated in $\mathcal{M}(TI)$ and $\mathcal{M}(TI) \models \beta$, and $\mathcal{M}(TI) \models \varphi$.

Finally, the language of motivation needs be amended once more to handle existential quantification.

existential motivation

Definition: A statement φ is *existentially motivated* in $\mathcal{M}(TI)$ if $\rho = \exists \mathsf{x}.\psi(\mathsf{x})$, $\rho \in CD$ or there is a causal rule $\alpha \wedge \beta \supset \rho \in T$ with α motivated in $\mathcal{M}(TI)$ and $\mathcal{M}(TI) \models \beta$, and φ is a skolemized existential specification of ρ, i.e. φ is what you obtain by substituting some unused skolem constant sk_i for each occurrence of x in ψ.[20]

An action is motivated—explained—whenever any of these conditions holds. To understand a theory instantiation, we simply minimize actions that are not explained according to these definitions.

We now say that a model is preferred if it has as few unmotivated actions as possible. A statement is unmotivated in $\mathcal{M}(TI)$ if it is not motivated in $\mathcal{M}(TI)$. Formally, we define the preference relation on models as follows:[21]

[19] Either ψ_1 or ψ_2—or both—can be empty. If both ψ_1 and ψ_2 are empty, semi-motivation reduces to strong motivation (if $\rho \in CD$) or to weak motivation (if $\alpha \wedge \beta \supset \rho \in T$).

[20] Semi-motivation and existential motivation of φ both depend on motivation of α and β. Since the times of α and β are guaranteed to be earlier than the times of φ, this is guaranteed not to lead to circularity.

[21] We had intended the previous version of preference, in [29, 35], to be equivalent to the current definition. Jonathan Amsterdam and Ramiro Guerreiro independently pointed out to us the error of our ways. The definition given here captures the intuitions intended by the previous version, is equivalent to it on all examples in the original paper, and corrects the non-transitivity of the original.

Definition: Let $unmot(\mathcal{M}(TI)) =$ *unmot*

$$\left\{ \mathsf{OCCURS(t, act)} \;\middle|\; \begin{array}{l} \mathcal{M}(TI) \models \mathsf{OCCURS(t, act)} \text{ and} \\ \mathsf{OCCURS(t, act)} \text{ is unmotivated in } \mathcal{M}(TI) \end{array} \right\}$$

$unmot(\mathcal{M}(TI))$ is the set of unmotivated actions in $\mathcal{M}(TI)$.

Then $\mathcal{M}_i(TI) \preceq \mathcal{M}_j(TI)$ ($\mathcal{M}_i$ is *preferable* to $\mathcal{M}_j$) if $unmot(\mathcal{M}_i(TI)) \subseteq unmot(\mathcal{M}_j(TI))$. $\preceq$

That is, $\mathcal{M}_i(TI)$ is preferable to $\mathcal{M}_j(TI)$ if "fewer" (subsetwise) unmotivated actions occur in $\mathcal{M}_i(TI)$. Note that such actions cannot be strongly motivated in $\mathcal{M}_j(TI)$; if an action is strongly motivated in one model, it is strongly motivated in *all* models.

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

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

Definition: $\mathcal{M}(TI)$ is a *preferred model* for TI if, for any model $\mathcal{M}'(TI) \preceq \mathcal{M}(TI)$, $\mathcal{M}'(TI) \approx \mathcal{M}(TI)$. *model*

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

We define the following sets: $\cap_{\mathcal{M}^\star}$

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

$\cup_{\mathcal{M}^\star} = \{\varphi \mid \exists \mathcal{M} \in \mathcal{M}^\star(TI), \mathcal{M} \models \varphi\}$—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. *projects*

Case II: φ is in $\cup_{\mathcal{M}^*(TI)}$. In this case, we say that φ is *consistent with TI*. However, if $\varphi \notin \cap_{\mathcal{M}^*(TI)}$, *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 $\mathsf{TIME}(\varphi)$ is later than the latest time point mentioned in *TI*, we say that *TI predicts* φ. *predicts*

5.3 Reasoning with MAT

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:

$$\begin{array}{l} \mathsf{HOLDS}(1, \mathsf{alive}) \\ \mathsf{OCCURS}(1, \mathsf{load}) \\ \mathsf{OCCURS}(5, \mathsf{shoot}) \end{array} \tag{31}$$

We have varied the statement slightly from (18), replacing HOLDS(1,loaded) with OCCURS(1,load) and delaying the shoot to 5 (as described on page 20). These changes do not affect the outcome, but allow us to better illustrate the effects of motivated action theory.

T contains causal rules for shoot, load, and unload, as well as the persistences for loaded and alive. The first causal rule is generated by statement (19); we have introduced the others because they will be useful below, but they do not effect the outcome of the original problem.

T: Causal Rules:

$$\begin{array}{l}
\forall \mathsf{t}.\mathsf{OCCURS}(\mathsf{t},\mathsf{shoot}) \wedge \mathsf{HOLDS}(\mathsf{t},\mathsf{loaded}) \supset \mathsf{HOLDS}(\mathsf{t}+1,\mathsf{not}(\mathsf{alive})) \\
\forall \mathsf{t}.\mathsf{OCCURS}(\mathsf{t},\mathsf{load}) \supset \mathsf{HOLDS}(\mathsf{t}+1,\mathsf{loaded}) \\
\forall \mathsf{t}.\mathsf{OCCURS}(\mathsf{t},\mathsf{shoot}) \supset \mathsf{HOLDS}(\mathsf{t}+1,\mathsf{not}(\mathsf{loaded})) \\
\forall \mathsf{t}.\mathsf{OCCURS}(\mathsf{t},\mathsf{unload}) \supset \mathsf{HOLDS}(\mathsf{t}+1,\mathsf{not}(\mathsf{loaded}))
\end{array} \tag{32}$$

Persistence Rules:

$$\begin{array}{l}
\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t},\mathsf{alive}) \\
\qquad \wedge((\neg\mathsf{OCCURS}(\mathsf{t},\mathsf{shoot})) \vee \ \mathsf{HOLDS}(\mathsf{t},\mathsf{not}(\mathsf{loaded}))) \\
\qquad\qquad \supset \mathsf{HOLDS}(\mathsf{t}+1,\mathsf{alive}) \\
\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t},\mathsf{not}(\mathsf{alive})) \supset \mathsf{HOLDS}(\mathsf{t}+1,\mathsf{not}(\mathsf{alive})) \\
\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t},\mathsf{loaded}) \\
\qquad \wedge\neg(\mathsf{OCCURS}(\mathsf{t},\mathsf{shoot})) \\
\qquad \wedge\neg(\mathsf{OCCURS}(\mathsf{t},\mathsf{unload})) \\
\qquad\qquad \supset \mathsf{HOLDS}(\mathsf{t}+1,\mathsf{loaded}) \\
\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t},\mathsf{not}(\mathsf{loaded})) \\
\qquad \wedge\neg\mathsf{OCCURS}(\mathsf{t},\mathsf{load}) \\
\qquad\qquad \supset \mathsf{HOLDS}(\mathsf{t}+1,\mathsf{not}(\mathsf{loaded}))
\end{array} \tag{33}$$

Note that these persistence rules will typically be derived from the causal rules and a closed-world assumption or compiled from CAUSES statements, as described above.

Consider the models of $TI = (31) \cup (32) \cup (33)$. Let $\mathcal{M}_1$ be the expected model, including HOLDS(5,loaded) and HOLDS(6,not(alive)); and let $\mathcal{M}_2$ be the unexpected model, where

HOLDS(5,not(loaded)), and therefore HOLDS(6,alive).[22] 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 HOLDS(1,alive), OCCURS(1,load), and OCCURS(5,shoot) are strongly motivated, since they are in CD. The fact HOLDS(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, HOLDS(6, not(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 $\mathcal{M}_1$. The only actions in $\mathcal{M}_1$, OCCURS(1,load) and OCCURS(5,shoot), are strongly motivated.

In contrast, $\mathcal{M}_2$ must entail another action. Since $\mathcal{M}_2 \models$ HOLDS(2, loaded) and also HOLDS(5,not(loaded)), something must defeat the persistence of loading. Therefore, $\mathcal{M}_2 \models$ OCCURS(t, unload) for some t $\in \{2,3,4\}$. (In fact, there are (at least) three such models, one in which the unload occurs at each of these times.) However, the occurrence of this unload action is not motivated: 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.

There is actually a third (class of) model(s), $\mathcal{M}_3(TI)$, which entails OCCURS(t,shoot), t $\in \{2,3,4\}$. Together with HOLDS(t,loaded), this entails HOLDS(t + 1, not(alive)). (If not HOLDS(t,loaded), then something must have caused the gun to become unloaded earlier, leading to another instance of the same problem.) Since not(alive) persists, this eventually gives us HOLDS(6,not(alive)). This model, however, contains an unmotivated action: OCCURS(t,shoot).

[22] There are of course many models of TI other than the two considered here....

In fact, it can be seen that in any preferred model of TI, HOLDS(5,loaded), and therefore HOLDS(6,not(alive)). That is because in any model where HOLDS(5,not(loaded)), a shoot or unload action must happen between time 2 and time 4, and such an action would be unmotivated. Since the facts HOLDS(5,loaded) and HOLDS(6,not(alive)) are in all preferred models of TI, TI projects these facts.

Causing Actions Nonetheless, preferring models in which the fewest possible unmotivated actions occur is not equivalent to preferring models in which the fewest possible actions occur. We can see this in cases where an action occurs on the right-hand side of a causal rule: in causal chains and in spontaneous actions.

Consider, e.g., the dominos example from section 2. T might include the causal rule

$$\begin{aligned} &\forall t, 0 < i < n. \\ &\qquad \mathsf{OCCURS}(t, \mathsf{fall}(\mathsf{domino}_i)) \\ &\qquad \wedge \neg \mathsf{OCCURS}(t, \mathsf{blockFall}(\mathsf{domino}_i)) \\ &\qquad\qquad \supset \mathsf{OCCURS}(t+1, \mathsf{fall}(\mathsf{domino}_{i+1})) \end{aligned} \tag{34}$$

Assume that the chronicle description contains $\mathsf{OCCURS}(1, \mathsf{fall}(\mathsf{domino}_1))$. Then, using our preference criterion, the theory instantiation projects

$$\forall 0 < i \leq n. \mathsf{OCCURS}(n, \mathsf{fall}(\mathsf{domino}_n)) \tag{35}$$

Minimizing actions would yield n minimal models, one agreeing with motivated action theory and

$\mathsf{n}-1$ additional models corresponding to the blockings of the $\mathsf{n}-1$ successive falls: For $1 \leq \mathsf{k} < \mathsf{n}$,

$$\begin{array}{c} \forall 1 \leq \mathsf{i} \leq \mathsf{k}.\mathsf{OCCURS}(\mathsf{i}, \mathsf{fall}(\mathsf{domino_i})) \\ \mathsf{OCCURS}(\mathsf{k}, \mathsf{blockFall}(\mathsf{domino_{k+1}})) \end{array} \tag{36}$$

and no other terms of the form OCCURS(t,act). 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.

A second non-equivalence of MAT and circumscribing action arises in the context of spontaneous actions. Consider, for example, the cookie thief of formula (29). Given *CD* containing HOLDS(1,in(cookies, kitchen)), MAT will yield two preferred models: $\mathcal{M}_1$, in which HOLDS(1,in(cookie-thief,kitchen)) and so OCCURS(2,cookie-theft); and $\mathcal{M}_2$, in which $\neg$HOLDS(1,in(cookie-thief,kitchen)) and so (assuming that T contains no other causal rules for cookie theft) $\neg$OCCURS(2,cookie-theft). MAT thus allows both possibilities, depending on the presence of the cookie thief. In contrast, minimizing actions prefers $\mathcal{M}_2$—no cookies stolen—unequivocally. In the authors' experience, this may be overly optimistic.

MAT here attempts to capture our intuition that some agents act autonomously and that their volitional actions can be motivated by their internal states. Such agents are represented by causal rules with no α-part—no positive occurrence terms. Similarly, occurrences such as sunrise can be represented by causal rules such as

$$\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t}, \mathsf{daybreak}) \supset \mathsf{OCCURS}(\mathsf{t}+1, \mathsf{sunrise}) \tag{37}$$

The elimination of such rules from T implies the absence of autonomously motivated agents and spontaneous—but caused—occurrences.

Backwards Projection We now show that our theory handles backward projection properly. As an example, consider the theory instantiation TI' consisting of the background theory (32) and (33), and the chronicle (21), in which HOLDS(6,alive). 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.[23] Motivated action 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 incomparable. 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. $\mathcal{M}'_4$ entails TI, 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:

$$\text{OCCURS}(2,\text{unload}) \vee \text{OCCURS}(3,\text{unload}) \vee \text{OCCURS}(4,\text{unload}) \tag{38}$$

which is exactly what we want.

5.4 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

[23] As we know, either an unload or a shoot will cause a gun to be unloaded. However, because we know that shooting will cause not(alive), that not(alive) persists forever, and that HOLDS(6,alive), all models for TI' must have an unload.

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 *occurrence kernel* is a pair $\langle A, B\rangle$, where A is a set of occurrence terms and B is a set of state terms. We define A complement as $\langle A, B\rangle$

$$\overline{A} \triangleq \{\neg \mathsf{OCCURS(t, act)} \mid \mathsf{OCCURS(t, act)} \notin A\}$$

and write $\langle A, B\rangle_{TI}$ for $TI \cup A \cup B \cup \overline{A}$. $\langle A, B\rangle_{TI}$

We say that an occurrence kernel $\langle A, B\rangle$ is *acceptable* for a theory instantiation TI if $\langle A, B\rangle_{TI} = TI \cup A \cup B \cup \overline{A}$ is consistent (whenever TI is). We say that $\langle A, B\rangle$ *supports* a statement φ if $\langle A, B\rangle_{TI} \vdash \varphi$.

An occurrence kernel thus determines the complete set of actions that do (A) and don't ($\overline{A}$) occur. If If B provides a value for every state at every time, then the "world" of $\langle A, B\rangle$ is completely determined—actions by A and $\overline{A}$, and state by B. However, we do not in general need a complete B. It is sufficient for the truth of HOLDS(t,state) to be derivable from $\langle A, B\rangle$. For example, the occurrence kernel

$$\begin{aligned} A &= \{\mathsf{OCCURS(1, load)}, \mathsf{OCCURS(5, shoot)}\} \\ B &= \{\mathsf{HOLDS(T_0, alive)}, \mathsf{HOLDS(T_0, not(loaded)}\} \end{aligned}$$

completely determines all state for $TI = (31) \cup (32) \cup (33)$. So, also, does

$$\begin{aligned} A &= \{\mathsf{OCCURS(1, load)}, \mathsf{OCCURS(2, unload)}, \mathsf{OCCURS(3, load)}, \mathsf{OCCURS(4, unload)}, \mathsf{OCCURS(5, shoot)}\} \\ B &= \{\mathsf{HOLDS(T_0, alive)}, \mathsf{HOLDS(T_0, loaded)}\} \end{aligned}$$

Here, we introduce the notion of $\mathsf{T_0}$, the *least time point*. We assume that $\mathsf{T_0}$ is a time point that precedes any time point mentioned in CD by some arbitrarily large (but finite) quantity. Further, we assume that $\forall \mathsf{act}.\neg\mathsf{OCCURS}(\mathsf{T_0}, \mathsf{act})$. Thus, from $\mathsf{HOLDS}(\mathsf{T_0}, \mathsf{loaded})$ and $\forall\ \mathsf{t},\ \mathsf{T_0} < \mathsf{t} < 1.\neg\ \mathsf{OCCURS}(\mathsf{t}, \mathsf{load})$ (which is in $\overline{A}$ for this $\langle A, B\rangle$), $\langle A, B\rangle_{TI}$ gives us $\forall \mathsf{t}, \mathsf{T_0} < \mathsf{t} \leq 1.\mathsf{HOLDS}(\mathsf{t}, \mathsf{loaded})$; similarly alive. *least time point*

Formally, if TI is a theory instantiation and $\langle A, B\rangle$ is an occurrence kernel acceptable for TI, then we say that $\langle A, B\rangle$ is *total* for TI if for every ground term $\varphi = \mathsf{HOLDS}(\mathsf{t}, \mathsf{state})$ or $\mathsf{OCCURS}(\mathsf{t}, \mathsf{act})$, *total*

$$\langle A, B\rangle_{TI} \vdash \varphi \qquad \text{or} \qquad \langle A, B\rangle_{TI} \vdash \neg\varphi$$

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 $\mathcal{OC}(TI)$ to be the set of occurrence kernels $\langle A, B\rangle$ that are both acceptable and total for TI. $\mathcal{OC}(TI)$

We now define the syntactic equivalent of motivation, the second order predicate $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$, recursively in terms of the first-order consequences of $\langle A, B\rangle_{TI}$. Condition 1 corresponds to strong motivation: it holds for necessary (monotonic) consequences of $CD \cup T = TI$, which are true in all models. Condition 2 covers the defeasible consequences of TI, i.e., those statements that are caused by a motivated action. These correspond to weakly motivated statements. Condition 3 parallels the definition of semi-motivation, in which a motivated disjunction means that (at least) one of the disjuncts is motivated. Finally, condition 4 mimics existential motivation, the motivation of *some* instantiation of a motivated existentially quantified statement. MOT

Definition: $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$ iff

1. $TI \vdash \varphi$, or

2. there exists in TI a causal rule of the form $\alpha \wedge \beta \supset \varphi$; $\mathsf{MOT}(\langle A, B\rangle, TI, \alpha)$; and $\langle A, B\rangle_{TI} \vdash \beta$, or

3. $\rho = \psi_1 \vee \varphi \vee \psi_2$; $\rho \in CD$ or a causal rule $\alpha \wedge \beta \supset \rho \in T$ with $\mathsf{MOT}(\langle A, B\rangle, TI, \alpha)$ and $\langle A, B\rangle_{TI} \vdash \beta$; and $\langle A, B\rangle_{TI} \vdash \varphi$, or

4. $\rho = \exists \mathsf{x}.\psi(\mathsf{x})$; $\rho \in CD$ or a causal rule $\alpha \wedge \beta \supset \rho \in T$ with $\mathsf{MOT}(\langle A, B\rangle, TI, \alpha)$ and $\langle A, B\rangle_{TI} \vdash \beta$; and φ is a skolemized existential specification of ρ, i.e. φ is what you obtain by substituting some unused skolem constant sk_i for each occurrence of x in ψ.

We next define the unmotivated actions in $\langle A, B\rangle$—the actions included in $\langle A, B\rangle_{TI}$ but not :

Definition: Let $\langle A, B\rangle$ be an acceptable and total occurrence kernel for TI. Then $\mathsf{unmot}(\langle A, B\rangle) =$

$$\left\{ \mathsf{OCCURS}(\mathsf{t}, \mathsf{act}) \;\middle|\; \begin{array}{c} \langle A, B\rangle_{TI} \vdash \mathsf{OCCURS}(\mathsf{t}, \mathsf{act}) \text{ and} \\ \neg \mathsf{MOT}(\langle A, B\rangle, TI, \mathsf{OCCURS}(\mathsf{t},\mathsf{act})) \end{array} \right\}$$

MOT induces a partial order on occurrence kernels. If $\langle A, B\rangle$ and $\langle A', B'\rangle$ are occurrence kernels in $\mathcal{OC}(TI)$, we say that $\langle A, B\rangle$ is preferred to $\langle A', B'\rangle$ ($\langle A, B\rangle \preceq \langle A', B'\rangle$) if $\mathsf{unmot}(\langle A, B\rangle) \subseteq \mathsf{unmot}(\langle A', B'\rangle)$. As with models, we call minimal elements under this ordering *preferred,* and call the set of these preferred occurrence kernels $\mathcal{OC}^\star(TI)$. We define $\cup_{\mathcal{OC}^\star(TI)}$ and $\cap_{\mathcal{OC}^\star(TI)}$ to be the union and intersection of statements in preferred occurrence kernels, respectively. $\cup_{\mathcal{OC}^\star(TI)}$ $\cap_{\mathcal{OC}^\star(TI)}$

Soundness and Completeness Below, we show that this definition of motivation is both sound and complete with respect to the semantic notion of motivation. We do this by demonstrating that any model has a corresponding (acceptable total) occurrence kernel which motivates the same actions—roughly its syntactic equivalent—and that any occurrence kernel has such a corresponding model. Thus, it follows that $\cup_{\mathcal{OC}^\star(TI)} = \cup_{\mathcal{M}^\star(TI)}$ and $\cap_{\mathcal{OC}^\star(TI)} = \cap_{\mathcal{M}^\star(TI)}$.

We begin by defining the mapping from models to corresponding occurrence kernels. For any model $\mathcal{M}$ and theory instantiation TI, this mapping will give us $\mathcal{OC}_{\mathcal{M},TI}$, an acceptable total occurrence kernel for TI which is essentially the syntactic equivalent of $\mathcal{M}$.

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

$$\begin{aligned} A &= \{\mathsf{OCCURS(t, act)} \mid \mathcal{M} \models \mathsf{OCCURS(t, act)}\} \\ B &= \{\mathsf{HOLDS}(\mathsf{T}_0, \mathsf{state}) \mid \mathcal{M} \models \mathsf{HOLDS}(\mathsf{T}_0, \mathsf{state})\} \\ &\cup \{\mathsf{HOLDS}(\mathsf{T}_0, \mathsf{not(state)}) \mid \mathcal{M} \models \mathsf{HOLDS}(\mathsf{T}_0, \mathsf{not(state)})\} \end{aligned}$$

That is, the occurrence kernel for a model agrees with that model on all actions—since every action is either in A or $\overline{A}$—and on enough state to make the occurrence kernel total for TI.

Formally, we have

Lemma 1.1 *Let TI be a theory instantiation; let $\mathcal{M}$ be a model for TI; and let $\mathcal{OC}_{\mathcal{M},TI} = \langle A, B\rangle$ be the occurrence kernel of $\mathcal{M}$ for TI. Then $\mathcal{OC}_{\mathcal{M},TI}$ is acceptable and total for TI.*

Proof: Proofs of all lemmas and theorems may be found in the appendix.

Since the occurrence kernel agrees with the model on all of the appropriate atomic formulae, semantic motivation in a model is equivalent to syntactic motivation in the occurrence kernel:

Theorem 1 *Let TI be a theory instantiation; let $\mathcal{M}$ be a model for TI; and let $\mathcal{OC}_{\mathcal{M},TI} = \langle A, B\rangle$ be the occurrence kernel of $\mathcal{M}$ for TI. Then $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$ iff φ is motivated in $\mathcal{M}$.*

In the other direction, we do not need to define a particular mapping from occurrence kernels to models; since models are necessarily "total," any model which entails $\langle A, B \rangle_{TI}$ is sufficient. This mapping from occurrence kernels to models is also motivation-preserving:

Theorem 2 *Let TI be a theory instantiation; let $\langle A, B \rangle$ be an occurrence kernel for TI; and let $\mathcal{M}$ be a model of $\langle A, B \rangle_{TI}$. Then φ is motivated in $\mathcal{M}$ iff* $\mathsf{MOT}(\langle A, B \rangle, TI, \varphi)$.

Given these correspondences, it is easy to see that the model-theoretic and proof-theoretic versions of motivation support the same conclusions:

Theorem 3 (Soundness and Completeness)

Let TI be a theory instantiation, with $\mathcal{M}^\star(TI)$ the set of preferred models for TI, and $\mathcal{OC}^\star(TI)$ the set of preferred occurrence kernels for TI. Then $\varphi \in \cup_{\mathcal{OC}^\star(TI)}$ iff $\varphi \in \cup_{\mathcal{M}^\star(TI)}$; $\varphi \in \cap_{\mathcal{OC}^\star(TI)}$ iff $\varphi \in \cap_{\mathcal{M}^\star(TI)}$.

5.5 Towards a Theory of 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 $\cap_{\mathcal{M}^\star(\mathrm{TI}_1)}$ 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 *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 by (32) and (33), and $CD_1 = \{\mathsf{HOLDS}(1, \mathsf{loaded}), \mathsf{HOLDS}(2, \neg\mathsf{loaded})\}$, and $TI_2 = T \cup CD_2$, where $CD_2 = CD_1 \cup \{\mathsf{OCCURS}(1, \mathsf{unload})\}$.

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 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 $\mathsf{OCCURS}(1, \mathsf{shoot}) \vee \mathsf{OCCURS}(1, \mathsf{unload})$.

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 problem, if we find out, after predicting $\mathsf{not(alive)}$, that $\mathsf{HOLDS(6, alive)}$. This is the sort of situation that demonstrates the nonmonotonicity of our logic, for TI_1 projects $\mathsf{HOLDS(6, \neg alive)}$, while $TI_2 \supset TI_1$ projects $\mathsf{HOLDS(6,alive)}$. 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 $\mathsf{Q} = \{\kappa \mid \kappa \in CD_2 \wedge \kappa \notin \cap_{\mathcal{M}^*(\mathrm{TI}_1)}\}$

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 with CD (21); let $TI_2 = T \cup CD_2$, where $CD_2 = CD_1 \cup \{\mathsf{HOLDS(6, alive)}\}$. The explanation of $\mathsf{HOLDS(6,alive)}$ 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.

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.

6 Discussion

The language that we used to describe temporal scenarios was adequate to the points that we wished to make here. However, most artificial intelligence applications will require a more realistic temporal ontology. Once we adopt such an ontology—for example, McDermott's full temporal logic [24]—the notion of causation that underlies motivated action theory will have to be revised. Although our central claim that *causation* is the underlying disambiguating principle of temporal reasoning still holds, a more sophisticated formalization of causation will ultimately be needed.

Morgenstern [28] has extended motivated action theory to provide the basis for an epistemic logic of action, called EMAT (Epistemic Motivated Action Theory). Most logics of action are not suitable for reasoning about *other agent's* knowledge and actions, either because they rely on complete enumeration of the actions taking place (*completeness*), or because they insist that *some* action—such as wait—must take place at every time point (*density*). Because motivated action theory is neither dense nor complete, it is possible to reason about periods during which some unknown actions may take place. This is critical to such reasoning processes as planning and plan recognition. EMAT explores these issues.

Amsterdam [5] suggests several improvements to motivated action theory. His disambiguating preference betters the notion of motivation in certain contexts, notably when performing backwards reasoning. Although motivated action theory correctly suggests that "something must have happened" in these scenarios, Amsterdam's *supported actions* allow more sophisticated reasoning about the nature of the intervening action. However, Amsterdam's supported action theory neither includes nor easily expands to cover phenomena such as causal chains.

When several legitimate possibilities exist, motivated action theory can only suggest a disjunction of these possibilities. For example, if a gun might have been unloaded at any time between its initial loading and its subsequent firing, motivated action theory remains uncommitted as to when

the unloading occurs. Further, if the gun might have been unloaded either by a wary gun-control activist, or by a Martian who happened to land nearby, motivated action theory can only assert that either of these scenarios is possible, in spite of the higher likelihood of the gun-control activist. To solve this problem, motivated action theory would ultimately need to be integrated with a theory of abductive inference.

Acknowledgements

This paper has benefitted immensely from discussions with Jonathan Amsterdam, Andrew Baker, Ken Bayse, Mark Boddy, Eugene Charniak, Ernie Davis, Tom Dean, Hector Geffner, Matt Ginsberg, Robert Goldman, Ramiro Guerreiro, Steve Hanks, Brian Haugh, Keiji Kanazawa, Vladimir Lifschitz, John McCarthy, Drew McDermott, Yoav Shoham, Solomon Shimony, and various anonymous referees.

References

[1] *Proceedings of the Fifth National Conference on Artificial Intelligence*, Philadelphia, Pennsylvania, August 1986. Morgan Kaufmann Publishers, Inc.

[2] *Proceedings of the Sixth National Conference on Artificial Intelligence*, Seattle, Washington, July 1987. Morgan Kaufmann Publishers, Inc.

[3] *Proceedings of the Seventh National Conference on Artificial Intelligence*, St. Paul, Minnesota, August 1988. Morgan Kaufmann Publishers, Inc.

[4] James F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154, 1984.

[5] Jonathan B. Amsterdam. Temporal reasoning and narrative conventions. In James F. Allen, Richard Fikes, and Erik Sandewall, editors, *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 15–21, Cambridge, Massachusetts, April 1991. Morgan Kaufmann Publishers, Inc.

[6] Andrew B. Baker and Matthew L. Ginsberg. Temporal projection and explanation. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 906–911, Detroit, Michigan, August 1989. Morgan Kaufmann Publishers, Inc.

[7] Frank M. Brown, editor. *The Frame Problem in Artificial Intelligence: Proceedings of the 1987 Workshop*, Lawrence, Kansas, April 1987. Morgan Kaufmann Publishers, Inc.

[8] Kenneth M. Ford and Patrick J. Hayes, editors. *Reasoning Agents in a Dynamic World: The Frame Problem*, volume 1 of *Advances in Human and Machine Cognition*, Greenwich, Connecticut, 1991. JAI Press. Also published as *International Journal of Expert Systems* 3(3–4).

[9] Michael Gelfond, Vladimir Lifschitz, and Arkady Rabinov. What are the limitations of the situation calculus? In *Working notes of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, 1991.

[10] Matthew L. Ginsberg and David E. Smith. Reasoning about action II: The qualification problem. *Artificial Intelligence*, 35:311–342, 1988.

[11] Steve Hanks and Drew V. McDermott. Temporal reasoning and default logics. Technical Report YALEU/CSD/RR 430, Department of Computer Science, Yale University, New Haven, Connecticut, October 1985.

[12] Steve Hanks and Drew V. McDermott. Default reasoning, nonmonotonic logics, and the frame problem. In *AAAI–86* [1], pages 328–333.

[13] Steve Hanks and Drew V. McDermott. Nonmonotonic logic and temporal projection. *Artificial Intelligence*, 33:379–412, 1987.

[14] Brian A. Haugh. Simple causal minimizations for temporal persistence and projection. In *AAAI–87* [2], pages 218–223.

[15] Patrick J. Hayes. The second naive physics manifesto. In Jerry Hobbs and Robert Moore, editors, *Formal Theories of the Commonsense World*, pages 1–36, Norwood, New Jersey, 1985. Ablex Publishing Co.

[16] Henry A. Kautz. The logic of persistence. In *AAAI–86* [1], pages 401–405.

[17] Vladimir A. Lifschitz. Pointwise circumscription: Preliminary report. In *AAAI–86* [1], pages 406–410.

[18] Vladimir A. Lifschitz. Formal theories of action: Preliminary report. In *AAAI–87* [2], pages 966–972.

[19] Vladimir A. Lifschitz and Arkady Rabinov. Miracles in formal theories of action. *Artificial Intelligence*, 38(2):225–237, March 1989. Research note.

[20] John M. McCarthy. Situations, actions, and causal laws. Technical Report 2, Stanford Artificial Intelligence Project, 1963.

[21] John M. McCarthy. Circumscription—A form of nonmonotonic reasoning. *Artificial Intelligence*, 13(1,2):27–39, 1980.

[22] John M. McCarthy. Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence*, 28(1):89–116, 1986.

[23] John M. McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4:463–502, 1969.

[24] Drew V. McDermott. A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6(2):101–155, 1982.

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

[26] Drew V. McDermott. AI, logic, and the frame problem. In Brown [7], pages 108–118.

[27] Drew V. McDermott and Jon Doyle. Non-monotonic logic I. *Artificial Intelligence*, 13(1,2):41–72, 1980.

[28] Leora Morgenstern. Knowledge and the frame problem. In Ford and Hayes [8].

[29] Leora Morgenstern and Lynn Andrea Stein. Why things go wrong: A formal theory of causal reasoning. In *AAAI–88* [3], pages 518–523.

[30] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1,2):81–132, 1980.

[31] Yoav Shoham. Chronological ignorance: Time, nonmonotonicity, necessity, and causal theories. In *AAAI–86* [1], pages 389–393.

[32] Yoav Shoham. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence*. MIT Press, Cambridge, Massachusetts, 1987.

[33] Yoav Shoham. Temporal logics in AI: Semantical and ontological considerations. *Artificial Intelligence*, 33:89–104, 1987.

[34] Lynn Andrea Stein. *Resolving Ambiguity in Nonmonotonic Reasoning.* PhD thesis, Department of Computer Science, Brown University, Providence, Rhode Island, 1990. Available as TR CS–90–18.

[35] Lynn Andrea Stein and Leora Morgenstern. Motivated action theory: A formal theory of causal reasoning. Technical Report CS–89–12, Department of Computer Science, Brown University, Providence, Rhode Island, March 1989.

A Proofs

Lemma 1.1 *Let TI be a theory instantiation; let $\mathcal{M}$ be a model for TI; and let $\mathcal{OC}_{\mathcal{M},TI} = \langle A, B\rangle$ be the occurrence kernel of $\mathcal{M}$ for TI. Then $\mathcal{OC}_{\mathcal{M},TI}$ is acceptable and total for TI.*

Proof:

Acceptable

$\mathcal{OC}_{\mathcal{M},TI} = \langle A, B\rangle$ is acceptable for TI iff $\langle A, B\rangle_{TI} = TI \cup A \cup B \cup \overline{A}$ is consistent (whenever TI is). In this case,

$$A = \{\mathsf{OCCURS(t, act)} \mid \mathcal{M} \models \mathsf{OCCURS(t, act)}\}$$

So

$$\overline{A} = \{\neg\mathsf{OCCURS(t, act)} \mid \mathsf{OCCURS(t, act)} \notin A\}$$

i.e.,

$$\overline{A} = \{\neg \mathsf{OCCURS}(\mathsf{t}, \mathsf{act}) \mid \mathcal{M} \not\models \mathsf{OCCURS}(\mathsf{t}, \mathsf{act})\}$$

And since $\mathcal{M} \models \varphi$ or $\mathcal{M} \models \neg\varphi$, $\forall\varphi$,

$$\overline{A} = \{\neg \mathsf{OCCURS}(\mathsf{t}, \mathsf{act}) \mid \mathcal{M} \models \neg\mathsf{OCCURS}(\mathsf{t}, \mathsf{act})\}$$

Also

$$\begin{aligned} B &= \{\mathsf{HOLDS}(\mathsf{T}_0, \mathsf{state}) \mid \mathcal{M} \models \mathsf{HOLDS}(\mathsf{T}_0, \mathsf{state})\} \\ &\cup \{\mathsf{HOLDS}(\mathsf{T}_0, \mathsf{not}(\mathsf{state})) \mid \mathcal{M} \models \mathsf{HOLDS}(\mathsf{T}_0, \mathsf{not}(\mathsf{state}))\} \end{aligned}$$

Finally, $\mathcal{M} \models TI$ by hypothesis. So $\mathcal{M} \models TI \cup A \cup B \cup \overline{A} = \langle A, B \rangle_{TI}$, and $\langle A, B \rangle_{TI}$ is consistent.

Total

$\mathcal{OC}_{\mathcal{M},TI} = \langle A, B \rangle$ is total for TI iff for every ground term $\varphi = \mathsf{HOLDS}(\mathsf{t}, \mathsf{state})$ or $\mathsf{OCCURS}(\mathsf{t}, \mathsf{act})$, $\langle A, B \rangle_{TI} \vdash \varphi$ or $\langle A, B \rangle_{TI} \vdash \neg\varphi$ Certainly, for $\varphi = \mathsf{OCCURS}(\mathsf{t}, \mathsf{act})$, either $\varphi \in A$ (and therefore $\langle A, B \rangle_{TI} \vdash \varphi$), or $\varphi \notin A$, (so $\neg\varphi \in \overline{A}$) so $\langle A, B \rangle_{TI} \vdash \neg\varphi$. Where $\varphi = \mathsf{HOLDS}(\mathsf{t}, \mathsf{state})$, the proof proceeds by induction on the number of time points from T_0 to t (which may be arbitrarily large but must be finite).

Base Case: Assume that $\mathsf{t} = \mathsf{T}_0 + 1$. Then by the definition of the least time point T_0, $\forall \mathsf{act}.\neg\mathsf{OCCURS}(\mathsf{T}_0, \mathsf{act})$. Thus, nothing can cause state to change from T_0 to t: if $\mathsf{HOLDS}(\mathsf{T}_0, \mathsf{state}) \in B$, $\langle A, B \rangle_{TI} \vdash \mathsf{HOLDS}(\mathsf{t}, \mathsf{state})$; if B contains

$\mathsf{HOLDS}(\mathsf{T_0}, \mathsf{not}(\mathsf{state}))$, $\langle A, B \rangle_{TI} \vdash \neg\mathsf{HOLDS}(\mathsf{t}, \mathsf{state})$.

Induction Hypothesis: Assume that $\langle A, B \rangle_{TI} \vdash \mathsf{HOLDS}(\mathsf{t}, \mathsf{state})$ or $\langle A, B \rangle_{TI} \vdash \neg\mathsf{HOLDS}(\mathsf{t}, \mathsf{state})$ whenever $\mathsf{t} - \mathsf{T_0} \leq k$, for some k.

Induction Step: Consider $\mathsf{t} = \mathsf{T_0} + k$. By the induction hypothesis either $\langle A, B \rangle_{TI} \vdash \mathsf{HOLDS}(\mathsf{t}-1, \mathsf{state})$ or $\langle A, B \rangle_{TI} \vdash \neg\mathsf{HOLDS}(\mathsf{t}-1, \mathsf{state})$. Assume without loss of generality that $\langle A, B \rangle_{TI} \vdash \mathsf{HOLDS}(\mathsf{t}-1, \mathsf{state})$.

Now the persistence rule for **state** looks something like

$$\begin{aligned} &\forall \mathsf{t}.\mathsf{HOLDS}(\mathsf{t}, \mathsf{state}) \\ &\quad \wedge \neg\ (cause_1) \\ &\qquad \vdots \\ &\quad \wedge \neg\ (cause_n) \\ &\qquad \supset \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{state}) \end{aligned}$$

where $cause_1 ... cause_n$ are

1. $\mathsf{OCCURS}(\mathsf{t}, \mathsf{act}) \wedge \mathsf{HOLDS}(\mathsf{t}, \mathsf{precond})$ whenever $\mathsf{CAUSES}(\mathsf{act}, \mathsf{precond}, \mathsf{not}(\mathsf{state}))$, or
2. $\alpha \wedge \beta$ whenever there is a causal rule $\alpha \wedge \beta \supset \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{not}(\mathsf{state}))$

We already have $\langle A, B \rangle_{TI} \vdash \mathsf{HOLDS}(\mathsf{t}, \mathsf{state})$. In addition, $cause_1 ... cause_n$ all involve times no later than **t**; so for each i, either $\langle A, B \rangle_{TI} \vdash cause_i$, or $\langle A, B \rangle_{TI} \vdash \neg cause_i$. If $\langle A, B \rangle_{TI} \vdash cause_i$, for some i, then (by the causal rule from which $cause_i$ is derived) $\mathsf{HOLDS}(\mathsf{t}+1, \mathsf{not}(\mathsf{state}))$; i.e., $\langle A, B \rangle_{TI} \vdash \neg\mathsf{HOLDS}(\mathsf{t}+1, \mathsf{state})$. If $\langle A, B \rangle_{TI} \nvdash cause_i$, for all i, then (by the induction hypothesis) $\langle A, B \rangle_{TI} \vdash \neg cause_i$ and so (by the persistence rule) $\langle A, B \rangle_{TI} \vdash \mathsf{HOLDS}(\mathsf{t}+1, \mathsf{state})$.

Lemma 1.2 *Let TI be a theory instantiation; let $\mathcal{M}$ be a model for TI; and let $\mathcal{OC}_{\mathcal{M},TI} = \langle A, B\rangle$ be the occurrence kernel of $\mathcal{M}$ for TI. Then $\langle A, B\rangle_{TI} \vdash \varphi$ iff $\mathcal{M} \models \varphi$*

Proof: First, we demonstrate that $\mathcal{M} \models \langle A, B\rangle_{TI}$:

$\mathcal{M} \models TI$.

$\mathcal{M} \models A$:

$$\mathcal{M} \models \{\mathsf{OCCURS(t, act)} \mid \mathcal{M} \models \mathsf{OCCURS(t, act)}\}$$

$\mathcal{M} \models B$:

$$\mathcal{M} \models \{\mathsf{HOLDS}(\mathsf{T}_0, \mathsf{state}) \mid \mathcal{M} \models \mathsf{HOLDS}(\mathsf{T}_0, \mathsf{state})\}$$
$$\cup \{\mathsf{HOLDS}(\mathsf{T}_0, \mathsf{not(state)}) \mid \mathcal{M} \models \mathsf{HOLDS}(\mathsf{T}_0, \mathsf{not(state)})\}$$

$\mathcal{M} \models \overline{A}$:

$$\mathcal{M} \models \{\neg\ \mathsf{OCCURS(t, act)} \mid \mathsf{OCCURS(t, act)} \notin A\}$$
i.e.
$$\{\neg\ \mathsf{OCCURS(t, act)} \mid \mathsf{OCCURS(t, act)} \notin A\}$$
or
$$\{\neg\ \mathsf{OCCURS(t, act)} \mid \mathcal{M} \not\models \mathsf{OCCURS(t, act)}\}$$

Since $\mathcal{M} \models \langle A, B\rangle_{TI}$, it follows that $\mathcal{M} \models \varphi$ whenever $\langle A, B\rangle_{TI} \vdash \varphi$. But by lemma 1.1, $\mathcal{OC}_{\mathcal{M},TI} = \langle A, B\rangle$ is total for TI, so (by the completeness of predicate calculus) $\langle A, B\rangle_{TI} \vdash \varphi$ whenever $\mathcal{M} \models \varphi$.

Theorem 1 *Let TI be a theory instantiation; let $\mathcal{M}$ be a model for TI; and let $\mathcal{OC}_{\mathcal{M},TI} = \langle A, B\rangle$ be the occurrence kernel of $\mathcal{M}$ for TI. Then* $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$ *iff φ is motivated in $\mathcal{M}$.*

Proof: (By temporal induction)

Base Case: Assume φ is of the form $\mathsf{HOLDS}(\mathsf{T_0}, \mathsf{state})$ or $\mathsf{OCCURS}(\mathsf{T_0}, \mathsf{act})$. There are two possibilities. If $TI \models \varphi$, then φ is motivated in $\mathcal{M}$ and also—since $TI \vdash \varphi$—$\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$. If $TI \not\models \varphi$, then φ cannot be motivated in $\mathcal{M}$ since no causal rule can have φ as its conclusion (by the definition of $\mathsf{T_0}$ as the least time point for TI) and no statement containing φ (as a disjunct or inside an existential quantifier) can appear in CD (again by definition of $\mathsf{T_0}$). But then also $\neg\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$.

Induction Hypothesis: Assume that φ is motivated in $\mathcal{M}$ iff $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$ whenever the time point mentioned in φ is strictly earlier than k.

Induction Step: Consider a statement φ with time k; i.e., $\varphi = HOLDS(k, state)$ or $\varphi = OCCURS(k, act)$. Assume first that φ is motivated in $\mathcal{M}$. Then there are four cases corresponding to the four types of motivation. If φ is strongly motivated, then $TI \models \varphi$, so $TI \vdash \varphi$, so $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$. If φ is weakly motivated, then there is a causal rule $\alpha \wedge \beta \supset \varphi \in T$, α is motivated in $\mathcal{M}$, and $\mathcal{M} \models \beta$. By the definition of a causal rule, the time of α is earlier than the time of φ, hence earlier than k, so $\mathsf{MOT}(\langle A, B\rangle, TI, \alpha)$; by lemma 1.2, $\langle A, B\rangle_{TI} \vdash \beta$; so $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$. If φ is semi- or existentially motivated, then either $\rho \in CD$ or ρ is the consequence of a causal rule with α motivated in $\mathcal{M}$ and $\mathcal{M} \models \beta$; then we have $\mathsf{MOT}(\langle A, B\rangle, TI, \alpha)$ by the induction hypothesis and $\langle A, B\rangle_{TI} \vdash \beta$ by lemma 1.2. Also, whenever $\mathcal{M} \models \varphi$, $\langle A, B\rangle_{TI} \vdash \varphi$ (by lemma 1.2). So whenever

φ is motivated in $\mathcal{M}$, $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$.

Conversely, if $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$, then φ is motivated in $\mathcal{M}$: If $TI \vdash \varphi$, then $TI \models \varphi$. If there is a causal rule of the form $\alpha \wedge \beta \supset \varphi \in T$ with $\mathsf{MOT}(\langle A, B\rangle, TI, \alpha)$ and $\langle A, B\rangle_{TI} \vdash \beta$, then α is motivated in $\mathcal{M}$ and $\mathcal{M} \models \beta$ (by the induction hypothesis and lemma 1.2, respectively). And whenever $\langle A, B\rangle_{TI} \vdash \varphi$, then (by lemma 1.2), $\mathcal{M} \models \varphi$. So whenever $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$, we also have that φ is motivated in $\mathcal{M}$.

Corollary 1.1 *Let TI be a theory instantiation; let $\mathcal{M}$ be a model for TI; and let $\mathcal{OC}_{\mathcal{M},TI} = \langle A, B\rangle$ be the occurrence kernel of $\mathcal{M}$ for TI. Then $unmot(\mathcal{M}) = \mathsf{unmot}(\mathcal{OC}_{\mathcal{M},TI})$*

Proof: This follows directly from theorem 1.

Lemma 2.1 *Let TI be a theory instantiation; let $\langle A, B\rangle$ be an occurrence kernel for TI; and let $\mathcal{M}$ be a model of $\langle A, B\rangle_{TI}$. Then $\mathcal{M} \models TI$.*

Proof: The proof of this is trivial: $\mathcal{M} \models \langle A, B\rangle_{TI}$ means $\mathcal{M} \models TI \cup A \cup B \cup \overline{A}$, so $\mathcal{M} \models TI$.

Lemma 2.2 *Let TI be a theory instantiation; let $\langle A, B\rangle$ be an occurrence kernel for TI; and let $\mathcal{M}$ be a model of $\langle A, B\rangle_{TI}$. Then $\mathcal{M} \models \varphi$ iff $\langle A, B\rangle_{TI} \vdash \varphi$.*

Proof: This follows directly from the soundness and completeness of predicate calculus.

Theorem 2 *Let TI be a theory instantiation; let* $\langle A, B\rangle$ *be an occurrence kernel for TI; and let* $\mathcal{M}$ *be a model of* $\langle A, B\rangle_{TI}$. *Then* φ *is motivated in* $\mathcal{M}$ *iff* $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$.

Proof: (By temporal induction)

Base Case: Assume φ is $\mathsf{HOLDS}(\mathsf{T_0}, \mathsf{state})$ or $\mathsf{OCCURS}(\mathsf{T_0}, \mathsf{act})$. There are two possibilities. If $TI \models \varphi$, then φ is motivated in $\mathcal{M}$ and also—since $TI \vdash \varphi$—$\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$. If $TI \not\models \varphi$, then φ cannot be motivated in $\mathcal{M}$ since no causal rule can have φ as its conclusion (by the definition of $\mathsf{T_0}$ as the least time point for TI) and no statement containing φ (as a disjunct or inside an existential quantifier) can appear in CD (again by definition of $\mathsf{T_0}$). But then also $\neg\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$.

Induction Hypothesis: Assume that φ is motivated in $\mathcal{M}$ iff $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$ whenever the time point mentioned in φ is strictly earlier than k.

Induction Step: Consider a statement φ with time k; i.e., $\varphi = HOLDS(k, state)$ or $\varphi = OCCURS(k, act)$. Assume first that φ is motivated in $\mathcal{M}$. Then there are four cases corresponding to the four types of motivation. If φ is strongly motivated, then $TI \models \varphi$, so $TI \vdash \varphi$, so $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$. If φ is weakly motivated, then there is a causal rule $\alpha \wedge \beta \supset \varphi \in T$, α is motivated in $\mathcal{M}$, and $\mathcal{M} \models \beta$. By the definition of a causal rule, the time of α is earlier than the time of φ, hence earlier than k, so $\mathsf{MOT}(\langle A, B\rangle, TI, \alpha)$; by lemma 2.2, $\langle A, B\rangle_{TI} \vdash \beta$; so $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$. If φ is semi- or existentially motivated, then either $\rho \in CD$ or ρ is the consequence of a causal rule with α motivated in $\mathcal{M}$ and $\mathcal{M} \models \beta$; then we have $\mathsf{MOT}(\langle A, B\rangle, TI, \alpha)$ by the induction hypothesis and $\langle A, B\rangle_{TI} \vdash \beta$ by lemma 2.2. Also, whenever $\mathcal{M} \models \varphi$, $\langle A, B\rangle_{TI} \vdash \varphi$ (by lemma 2.2). So whenever

φ is motivated in $\mathcal{M}$, $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$.

Conversely, if $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$, then φ is motivated in $\mathcal{M}$: If $TI \vdash \varphi$, then $TI \models \varphi$. If there is a causal rule of the form $\alpha \wedge \beta \supset \varphi \in T$ with $\mathsf{MOT}(\langle A, B\rangle, TI, \alpha)$ and $\langle A, B\rangle_{TI} \vdash \beta$, then α is motivated in $\mathcal{M}$ and $\mathcal{M} \models \beta$ (by the induction hypothesis and lemma 2.2, respectively). And whenever $\langle A, B\rangle_{TI} \vdash \varphi$, then (by lemma 2.2), $\mathcal{M} \models \varphi$. So whenever $\mathsf{MOT}(\langle A, B\rangle, TI, \varphi)$, we also have that φ is motivated in $\mathcal{M}$.

Corollary 2.1 *Let TI be a theory instantiation; let $\langle A, B\rangle$ be an occurrence kernel for TI; and let $\mathcal{M}$ be a model of $\langle A, B\rangle_{TI}$. Then* $\mathsf{unmot}(\langle A, B\rangle) = unmot(\mathcal{M})$.

Proof: This follows directly from theorem 2.

Lemma 3.1 *Let TI be a theory instantiation; let $\langle A, B\rangle$ be an occurrence kernel for TI; and let $\mathcal{M}$ be a model of $\langle A, B\rangle_{TI}$. Then $\langle A, B\rangle$ is a preferred occurrence kernel iff $\mathcal{M}$ is a preferred model.*

Proof: Assume that $\langle A, B\rangle$ is a preferred occurrence kernel of TI, but $\mathcal{M}$ is not a preferred model of TI. Then there is some model $\mathcal{M}'$ such that $unmot(\mathcal{M}') \subset unmot(\mathcal{M})$. Consider $\mathcal{OC}_{\mathcal{M}', TI}$ the occurrence kernel of $\mathcal{M}'$ for TI. $\mathsf{unmot}(\mathcal{OC}_{\mathcal{M}', TI}) = unmot(\mathcal{M}')$ by corollary 1.1. By corollary 2.1, $\mathsf{unmot}(\langle A, B\rangle) = unmot(\mathcal{M})$. But $\mathsf{unmot}(\langle A, B\rangle) = unmot(\mathcal{M}) \subset unmot(\mathcal{M}') = \mathcal{OC}_{\mathcal{M}', TI}$, and $\langle A, B\rangle$ is not preferred; contradiction!

Now assume that $\langle A, B\rangle$ is not preferred, i.e. $\exists \langle A', B'\rangle.\mathsf{unmot}(\langle A', B'\rangle) \subset \mathsf{unmot}(\langle A, B\rangle)$. Consider $\mathcal{M}'$, a model of $\langle A', B'\rangle$. $unmot(\langle A', B'\rangle) = unmot(\mathcal{M}')$ by corollary 2.1; similarly, $\mathsf{unmot}(\langle A, B\rangle) = unmot(\mathcal{M})$; so $unmot(\mathcal{M}') \subset unmot(\mathcal{M})$, and $\mathcal{M}$ is not preferred.

Lemma 3.2 *Let TI be a theory instantiation; let $\mathcal{M}$ be a model for TI; and let $\mathcal{OC}_{\mathcal{M},TI} = \langle A, B\rangle$ be the occurrence kernel of $\mathcal{M}$ for TI. Then $\mathcal{OC}_{\mathcal{M},TI}$ is a preferred occurrence kernel iff $\mathcal{M}$ is a preferred model.*

Proof: Since $\mathcal{M}$ is a model for its occurrence kernel, this is simply a special case of the previous lemma.

Theorem 3 (Soundness and Completeness)

Let TI be a theory instantiation, with $\mathcal{M}^*(TI)$ the set of preferred models for TI, and $\mathcal{OC}^*(TI)$ the set of preferred occurrence kernels for TI. Then $\varphi \in \cup_{\mathcal{OC}^*(TI)}$ iff $\varphi \in \cup_{\mathcal{M}^*(TI)}$; $\varphi \in \cap_{\mathcal{OC}^*(TI)}$ iff $\varphi \in \cap_{\mathcal{M}^*(TI)}$.

Proof: If $\varphi \in \cup_{\mathcal{OC}^*(TI)}$, then there is a preferred occurrence kernel $\langle A, B\rangle$ of TI such that $\langle A, B\rangle_{TI} \vdash \varphi$. Consider $\mathcal{M}$, a model of $\langle A, B\rangle_{TI}$: by lemma 2.2, $\mathcal{M} \models \varphi$; by lemma 3.1, $\mathcal{M}$ is a preferred model of TI. So $\varphi \in \cup_{\mathcal{M}^*(TI)}$.

Conversely, if $\varphi \in \cup_{\mathcal{M}^*(TI)}$, then there is a preferred model $\mathcal{M}(TI)$ such that $\mathcal{M}(TI) \models \varphi$. Consider $\mathcal{OC}_{\mathcal{M},TI}$, the occurrence kernel of $\mathcal{M}$ for TI: by lemma 1.2, $\mathcal{OC}_{\mathcal{M},TI} \vdash \varphi$; by lemma 3.2, $\mathcal{OC}_{\mathcal{M},TI}$ is a preferred occurrence kernel of TI. So $\varphi \in \cup_{\mathcal{OC}^*(TI)}$.

If $\varphi \in \cap_{\mathcal{OC}^*(TI)}$, then every preferred occurrence kernel of TI supports φ. Since preferred occurrence kernels are total, this means that no occurrence kernel supports $\neg\varphi$; i.e., $\neg\varphi \notin \cup_{\mathcal{OC}^*(TI)}$. But then $\neg\varphi \notin \cup_{\mathcal{M}^*(TI)}$, either, so (since every model entails either φ or $\neg\varphi$) $\varphi \in \cap_{\mathcal{M}^*(TI)}$.

Similarly, $\varphi \in \cap_{\mathcal{M}^*(TI)}$ means that $\neg\varphi \notin \cup_{\mathcal{M}^*(TI)}$, so $\neg\varphi \notin \cup_{\mathcal{OC}^*(TI)}$, so $\varphi \in \cap_{\mathcal{OC}^*(TI)}$.