

Knowledge and the Frame Problem

Leora Morgenstern

IBM T. J. Watson Research Center

P.O. Box 704

Yorktown Heights, NY 10598

(914)784-7151

leora@watson.ibm.com

1 Abstract

The frame problem, which arises in any reasonably expressive theory of action, shows up in a strengthened form in integrated theories of knowledge and action. In this paper, we present two instances of this strengthened frame problem: the Third Agent Frame Problem and the Vicarious Agent Frame Problem. We show that these problems cannot be solved at all by using frame axioms, and that most non-monotonic temporal logics are likewise inadequate for handling these problems. Finally, we show how an existing non-monotonic temporal logic, Motivated Action Theory, can be extended to handle both problems.

2 Introduction

It has long been recognized that agents who operate in complex environments must be able to reason about the interactions between knowledge and action ([McCarthy and Hayes, 1969], [Moore, 1980], [Konolige, 1982]). Agents must reason about the knowledge they need in order to perform an action, how actions that they perform affect their knowledge, and how they can plan to achieve their goals with only partial knowledge. The bulk of the research in this area has focussed upon the first of these issues, more commonly known as the *knowledge preconditions problem*. Specifically, researchers have sought to formalize in

a concise manner the principles which agents use to reason about their abilities to perform actions. Most of this work has emphasized the development of epistemic logics, but has been content to remain within existing temporal logics, such as the situation calculus (used by [McCarthy and Hayes, 1969], [Moore, 1980], and [Konolige, 1982]) or interval-based temporal logics such as McDermott’s [1982] (used by [Morgenstern, 1988]).

Little work has been done, however, on the ways in which one would have to modify a theory of action in order to successfully integrate it with a theory of knowledge. In particular, it has been an implicit assumption of the researchers who have worked out detailed problems in this area ([Moore, 1980], [Morgenstern, 1988]) that the frame problem¹, or temporal projection problem, that arises in virtually any reasonably expressive theory of action carries over in the same form to an integrated theory of knowledge and action. It has been believed, as a consequence, that it can be solved by the same means as the original frame problem is solved.

In this paper, we show that two particularly virulent variants of the frame problem arise when we integrate a theory of knowledge with a theory of action that allows for multiple agents. The first of these problems, the Third Agent Frame Problem, arises when an agent constructing a plan cannot be sure that a minor participant in the plan will know enough to play his part in the plan when his time to act comes around. The second of these problems, the Vicarious Planning Frame Problem, arises when the agent who constructs the plan *delegates* a part of the plan, and cannot predict what the world will be like once the delegated portion of the plan has been completed.

We will discuss both of these problems in detail in the next section. In Section 4, we will show that the standard solution to the frame problem — frame axioms — are of absolutely no help in solving these frame problems. We also show that one cannot merely add a “dose of non-monotonicity” [McDermott, 1982] to solve these problems; that, in fact, most of the non-monotonic logics that have been developed in recent years (e.g., [Lifschitz, 1987], [Haugh, 1987], [Kautz, 1986], [Baker and Ginsberg, 1989]) are totally unsuited to the task. Finally, we will show how a more expressive non-monotonic temporal logic, Motivated Action Theory [Morgenstern and Stein, 1988], can be extended to solve the Third Agent

¹See footnote 2

Frame Problem and the Vicarious Planning Frame Problem.

3 The Variant Frame Problems

The temporal projection problem, or as it is more commonly called, the frame problem,² is essentially the problem of determining what facts about the world stay the same when an action is performed. It is presumably the case that for most actions, most facts about the world stay the same after that action has been performed. For example, after an agent moves block X onto block Y, the blocks are the same color, the furniture in the room is still in place, the president is still in office, and so on. The earliest approach to this problem [McCarthy and Hayes, 1969] called for adding axioms to a particular theory specifying how actions *did not* change facts. Thus, in the example above, one might add axioms stating that moving a block does not change its color or the location of any furniture in the room.

Frame axioms allow a system with complete knowledge of all the facts in the theory to make the desired inferences. For example, consider a system with a frame axiom that states that if block X has color C in situation z1, and situation z2 results from moving block X in z1, then X has color C in z2. Now, suppose that block Q1 is moved to block Q2 between situations s1 and s2, and suppose that block Q1 is red at s1. Then the system could infer that block Q1 is red at s2. In this sense, frame axioms can be said to offer a solution of sorts to the temporal projection problem. It is without doubt a problematic solution as [McCarthy and Hayes, 1969], [Hayes, 1977], and [McDermott, 1984] have noted. Firstly, a large number of frame axioms are needed for even simple theories, making such theories difficult to formalize and intractable. Secondly, as we will discuss in more detail in the next section, frame axioms do not seem to parallel the sort of commonsense inferences that

²This common usage is in fact incorrect. McCarthy and Hayes [1969] discussed, but did not give a name to, the temporal projection problem within the context of the situation calculus. They suggested adding frame axioms to solve this problem, and designated as the *frame problem* the problem arising from the proliferation of frame axioms that were needed to handle the temporal projection problem. The frame problem, thus, as it was originally formulated, arises only within the situation calculus. Since the common usage of the term *frame problem* refers to the temporal projection problem, and since McCarthy and Hayes originally used the term with this meaning in mind [personal communication], we will, in accordance with the adage that meaning is use, use the term in that way throughout this paper.

human beings do when they engage in temporal projection. Finally, frame axioms are often false in a system that allows for concurrent actions - what is to prevent someone from spray painting block Q1 as it is being placed on block Q2? Thus, they are typically suited only for systems that do not allow concurrent actions. Nonetheless, frame axioms have gained *de facto* acceptance among many AI researchers. In simple cases, they do allow the desired inferences to go through, and often, that is all that is needed. It's ugly, but it works.

3.1 Integrated Theory of Knowledge and Action: Simple Cases

Frame axioms, even if not ideal, are adequate for handling simple instances of the temporal projection problem in an integrated theory of knowledge and action. Consider a system that allows for incomplete and partial knowledge and that does not allow concurrent actions, and consider the following problem:

Example 1:

Bill wants to open a safe. To open the safe, he needs to know the combination of the safe. He does not know the combination, but knows that his friend Susan does. Assume that friendly agents cooperate and give information when requested to do so. To further simplify the problem, assume that in this particular problem situation, concurrent actions are not allowed. How can Bill plan to open the safe?

This seemingly simple problem is actually based on a rather complex theory of knowledge preconditions for actions and plans. A complete description of such a theory and a complete axiomatization of the above problem within a robust temporal ontology is given in [Morgenstern, 1988]. Since such a theory is not the focus of the present paper, we present below a simplified (and sometimes approximate) version of the theory. We will be working in a simple epistemic temporal logic, \mathcal{E} . \mathcal{E} is an instance of the first order predicate calculus. Statements in \mathcal{E} may be temporal or atemporal. Examples of atemporal statements are theorems of arithmetic or universal truths. Examples of temporal statements are

Lives(6,Edward, NYC)

(Edward lives in New York City at time 6)

and

$\exists p$ (Person(p) and Taller-than(4,Amy, p))
 (Amy is taller than some person at time 4).

In limited situations, we allow ourselves to “factor out” the time argument by introducing some syntactic sugar. If a temporal statement x is of the form

$[Q arg_1] \dots [Q arg_n] P(t, arg_1, \dots, arg_t)$

where Q stands for \forall or \exists , we may write

True($t, [Q arg_1] \dots [Q arg_n] P(arg_1, \dots, arg_t)$)

Thus, for example, the above examples could be rewritten as

True(6,Lives(Edward, NYC))

and

True(4, $\exists p$ (Person(p) and Taller-than(Amy, p)))

No meaning should be read into the True predicate; it is merely syntactic sugar.

The temporal logic used is rather simple. Time points are assumed to be isomorphic to the integers. Any two time points define an *interval*; however, since time points are isomorphic to the integers, any finite interval will contain only a finite number of time points. Following [McDermott, 1982], we define an action to be the set of intervals in which it occurs. For example, the action of moving block X to block Y is the set of all intervals in which block X is moved to block Y. An *action* is distinguished from an *event*: an event is an action restricted to a particular agent. We speak of the *action* of riding a bicycle, but the *event* of Hilary riding a bicycle. If a is an agent and act is an action, Do(a, act) is the event of a performing act.

In virtually all (implemented or non-implemented) temporal logics, actions have been described using functions. This is true whether they have been viewed ontologically as functions on states or as sets of intervals. For example, a common way of representing the action of moving block X to block Y is Move(X,Y). Often, however, functions are not adequately descriptive. For example, there is no natural, straightforward way to describe

the action of moving any red block onto any blue block using functions ³. Using interval descriptions, however, we can easily describe this action as

$$\{i \mid \exists a,x,y \text{ Red}(x) \text{ and Blue}(y) \text{ and Do}(a,\text{Move}(a,b),i)\} . \text{ } ^4$$

Any action which can be described using functions can be described using intervals, though the converse is not true. In this paper, we will use functional descriptions whenever possible (they are in general less cumbersome and easier to understand), and use interval descriptions when they are needed.

Actions are assumed to take unit time — i.e., the *length* of the action is one — unless otherwise specified. If $ev = do(a,act)$, then $length(ev) = length(act)$. Plans are defined recursively: an event is a plan, and if $pln1$ and $pln2$ are plans and c is a condition (i.e. $True(t,c)$ is a well-formed sentence), $sequence(pln1,pln2)$, $concurrent(pln1,pln2)$, $ifthenelse(c,pln1,pln2)$, and $while(c,pln1)$ are all plans. The *length* of a plan, $length(pln)$ is defined as follows: If pln is of the form $sequence(pln1,pln2)$, $length(pln) = length(pln1) + length(pln2)$; if pln is of the form $ifthenelse(c,pln1,pln2)$, $length(pln) = length(pln1)$ if c is true and $length(pln2)$ if c is false; if pln is of the form $concurrent(pln1,pln2)$, $length(pln) = \max(pln1,pln2)$. If pln is of the form $while(c,pln1)$, we leave $length(pln)$ undefined. In this paper, we will be focussing on the sequence construct, and ignoring the branch, loop, and concurrency constructs.

A plan need not consist of actions performed by a single agent. In fact, most plans of interest involve many active agents. For example, my plan to get my car fixed involves my driving my car to my mechanic, the mechanic inspecting the car, diagnosing the problem, and fixing the car, and my paying the mechanic. Often, plans can be described functionally.

³There are, of course, any number of *unnatural* ways to describe this action. We could sort our logic so that variables beginning with g range over red blocks and variables beginning with h range over blue blocks. Then we could describe the action as $Move(g,h)$. Alternatively, we could define a new action called $MoveRB$, semantically equivalent to the set of all intervals in which red blocks are moved to blue blocks, and then describe the action as $MoveRB(x,y)$. Such schemes, while they would work for the case at hand, are clearly ad hoc kludges, and are not desirable strategies for the general case.

⁴Note, above, the overloading of the symbol Do , which is used both as a function with two arguments, an agent and an action, and a predicate with three arguments, an agent, an action, and an interval.

We have $Do(a,act,i) \Leftrightarrow i \in Do(a,act)$.

My plan above can be described as

```
sequence( do(I,drive(Car22,Mechanic55)),
          do(Mechanic55,inspect(Car22)),
          do(Mechanic55,diagnose(Car22)),
          do(Mechanic55,fix(Car22)),
          do(I,pay(Mechanic55,price(repairs(Car22))))))
```

Other less specified plans cannot be described functionally. For example, my plan to get my car fixed by any mechanic who lives in NYC and is reasonably priced could not be described functionally (in a natural and straightforward manner), but could be described using intervals, as:

```
{i1,i2,i3, i4,i5 | ∃ m In(m, NYC) ∧ Reasonably-Priced(m) ∧
  sequence( do(I,drive(Car22,m),i1),
            do(m,inspect(Car22),i2),
            do(m,diagnose(Car22),i3),
            do(m,fix(Car22),i4),
            do(I,pay(m,price(repairs(Car22))),i5))}
```

Knowledge is represented using the predicate *Know*. *Know* takes two arguments, an agent and a term representing a sentence.⁵ Strictly speaking, we should delimit our sentences with quotation marks to demonstrate this. In fact, many of the predicates that we use in this paper, such as *Bel*, *Know-what-is*, and *Can-perform* are opaque, syntactic predicates which take terms *representing* sentences, actions, or objects. The arguments of these predicates should likewise be delimited by quotation marks. Throughout this paper, however, we will be suppressing quotation marks. We do this because (1) quotation marks and quasi-quotation marks would make the text quite difficult to read, and (2) the quotation construct is not relevant to the research foci of this paper. The reader is referred to [Morgenstern, 1988] for a full treatment of quotation and quantifying into epistemic contexts.

⁵We use a syntactic logic of knowledge rather than a modal logic of knowledge due to its greater expressivity. In particular, when using a modal logic, one cannot express statements such as "Bill knows that Susan knows something he doesn't know" or "Ted knows more than Frank." Such expressivity is essential for a robust theory of partial knowledge.

Determining when an agent can be said to know what a particular object is is of particular importance in a theory of knowledge. Philosophers and linguists alike have discussed this subject; see, e.g., [Moore, 1980]. In general, the answer seems to be context-dependent. In many cases, however, it makes sense to say that an agent knows what a particular object is, or who someone is, if he knows a *standard identifier* — i.e., a name or number, for that object or person. For example, I can be said to know what Alice’s phone number is if I know that her number is 777-9999; likewise I know who Rick’s ophthalmologist is if I know that Dr. Rachel Feelwell is Rick’s ophthalmologist. We introduce the predicate *Know-what-is* to formalize this concept. The predicate takes two terms, aside from the temporal argument: an agent and a string representing some term. We have the following definition: (we again suppress quotation marks here to improve readability)

Definition 1

$\text{True}(t, \text{Know-what-is}(a, \text{trm}))$

$\Leftrightarrow \exists p \text{ True}(t, \text{Know}(a, \text{Standard-id}(p) \text{ and } \text{Equal}(p, \text{trm})))$

Given this language our two boundary conditions can be stated as follows:

Boundary Conditions 3.1

1. $\text{True}(1, \text{Know}(\text{Bill}, (\text{True}(1, \text{Know-what-is}(\text{Susan}, \text{comb}(\text{Sf1}, 1))))))$
 (Bill knows at time 1 that Susan at time 1 knows what the combination of the safe is. Note that the combination of the safe is a fluent; i.e. it can change over time)
2. $\text{True}(1, \text{Friendly}(\text{Bill}, \text{Susan}))$
 (Bill and Susan are friendly at time 1)

The axioms for the simplified theory of knowledge preconditions and plans are given in Appendix A. Informally, these axioms state that friendly agents will do what they are asked, if they can, and will plan to do what they are asked, if at all possible; that an agent can dial a safe as long as he knows the combination number and is at the safe; that friendly agents can give over information only if they know it to be true. Also given are the knowledge preconditions for the execution of plans. Essentially, an agent can execute a plan if he knows that he will be able to perform his parts in the plan at the proper times, and

can predict that the other parts of the plan will occur at the proper times. For example, consider my plan, above, for getting my car fixed. I can execute this plan if I know that I will be able to drive my car to the mechanic, and will, when the time comes, be able to pay the mechanic for the repairs to my car, and if, in addition, I can predict that the mechanic will indeed examine my car, diagnose the problem, and repair the car. It is not enough for me to know that the mechanic is *capable* of examining, diagnosing, and repairing my car; unless he *does* these actions, my plan will not work.

The theory as we have thus far informally described it, and as it is formally described in Appendix A, is not quite sufficient, however to solve Example 1. Bill should be able to reason as follows: if I ask Susan for the combination of the safe, say between times 1 and 2, then Susan will give me the information, say between times 2 and 3, and at time 3, I will know how to dial the combination of the safe.

In order for Bill to reason that Susan will give over the information at time 2, he must know, among other things, that at time 2 Susan knows the combination of the safe at that time. Now, he knows that Susan knows at time 1 what the combination of the safe is. But how can he know that Susan knows at time 2 what the combination of the safe is? To make that inference go through, we must add the assumptions (1) that all agents know all frame axioms, and (2) that the speaker and hearer of a communicative event always know when the action has taken place. Then Bill will know that at time 2, Susan will still know what the combination of the safe at time 1 is. Of course, at time 2, Susan needs to know what the combination of the safe at time 2 is. Since Bill can reason that Susan can reason with frame axioms, he can conclude that Susan can figure out at time 2 that the combination at time 2 is exactly what it was at time 1. We thus have the following theorem:

Theorem 1

$$\begin{aligned} &\text{True}(1, \text{Occurs}(\text{sequence} \\ &\quad (\text{do}(\text{Bill}, \text{request}(\text{Susan}, \text{tell-what-is}(\text{comb}(\text{Sf1}, 2))))), \\ &\quad \text{do}(\text{Susan}, \text{tell-what-is}(\text{comb}(\text{Sf1}, 2)))))) \\ &\quad \Rightarrow \\ &\text{True}(3, \text{Know-how-to-perform}(\text{Bill}, \text{dial}(\text{comb}(\text{Sf1}, 3)))). \end{aligned}$$

Both assumptions (1) and (2) are easy to add to the theory. In a modal logic of knowledge, such as Moore [1980], agents automatically know all axioms, including frame axioms. In other logics, it is a simple matter to posit that agents know frame axioms. It is likewise quite natural to assume that agents always know when they are speaking or listening.

3.2 The Third Agent Frame Problem

As soon as we introduce a slightly more complicated version of this problem, it becomes apparent that the theory we have set up is not sufficiently powerful. We consider now a case that superficially appears almost identical to Example 1. The only difference seems to be that the ignorant agent needs to make two requests (as opposed to one request) to obtain the information that he needs. This difference turns out to be very significant; as we shall see, such a problem cannot be solved within \mathcal{E} as it has so far been presented.

Example 2: Let us assume that Bill at time 1 wants to open the safe. His knowledge in this case, however, is different from, and weaker than, his knowledge in Example 1. Bill knows, not that Susan knows the combination of the safe, but that Susan knows some friend of Bill, and that this friend of Bill knows the combination of the safe.

Thus, we have the following initial conditions:

Boundary Conditions 3.2

1. $\text{True}(1, \text{Know}(\text{Bill},$
 $\quad \text{True}(1, \exists a \text{ Know}(\text{Susan},$
 $\quad \quad \text{True}(1, \text{Friendly}(a, \text{Bill}, 1) \wedge \text{Know-what-is}(a, \text{comb}(\text{Sf}1, 1))))))$
2. $\text{True}(1, \text{Friendly}(\text{Bill}, \text{Susan}))$

At first glance, this problem just seems like more of the same of Example 1. Bill should be able to convince himself that the following four-step plan will work:

- Bill asks Susan for the name of the third agent
- Susan tells Bill the name of the third agent
- Bill asks the third agent for the combination

- the third agent tells Bill the combination

Unfortunately, Bill cannot prove that this plan will work. Bill knows that the third agent knows the combination at time 1, and he knows that the combination will remain the same from time 1 to time 3. However, he cannot prove that the third agent will know at time 3 that he still knows the combination of the safe at that time. This is true even if it is assumed that all agents, and in particular the third agent, know all the frame axioms. The third agent is not necessarily aware of the actions that have taken place between times 1 and 3, and therefore cannot apply the frame axioms.

This problem, which we shall call the Third Agent Frame Problem, effectively prevents Bill from proving that his perfectly reasonable plan will work. More generally, it prevents any agent from knowing that he can successfully execute a plan if that plan involves a chain of requests and it cannot be assumed that agents in the plan know that the previous chain of events has occurred. Such plans are obviously quite common in real life: consider, for example, all the times one plans to ask for a referral of an expert, and subsequently ask the expert for an opinion. The expert will typically not be aware of the referral, but this in no way affects her expertise. If I ask a friend for the name of a good orthopedist, and subsequently visit the orthopedist, the fact of my asking my friend for a recommendation should be irrelevant to the orthopedist's ability to set a fracture. Clearly a theory of knowledge and planning that does not recognize this truth and that cannot handle such conceptually simple plans is severely limited.

In Section 3, we will discuss in detail why the Third Agent Frame Problem arises. In the meantime, we note that the existence of the problem is not due to an undersupply of frame axioms. No amount of extra frame axioms will allow Bill to prove that the third agent can prove that he still knows the combination of the safe at time 3. It is not the frame axioms or even the *knowledge* of the frame axioms that is lacking, it is knowledge of the events that have occurred that is lacking. Likewise, the existence of the problem is not due to the possibility that more than one action can happen at a time. This problem arises even in a system that does not allow for concurrent actions. Thus, even the drastic simplifications employed by most AI researchers are not sufficient for this problem.

3.3 The Vicarious Planning Frame Problem

The fact that the active agents in a plan are not necessarily aware of all the actions in the plan was the underlying cause of the Third Agent Frame Problem; it is likewise the underlying cause of the next variant frame problem, discussed below. Here, it is the most active agent in the plan who is not aware of all the actions in the plan, since he delegates part of the plan to other agents.

Example 3: We consider another planning situation, quite similar to the one discussed above. Bill wants to open the safe (at time 1). He knows that he is located near the safe and that he is thus physically capable of opening the safe, but he still needs to know the combination. Again, he has some partial information: he knows that his friend Susan knows someone who knows the combination. In this case, he does not know that the third agent is a friend of his; rather, he knows that the third agent is a friend of Susan. Formally, we state these conditions as follows:

Boundary Conditions 3.3

1. $\text{True}(1, \text{At}(\text{Bill}, \text{Sf1}))$
2. $\text{True}(1, \text{Friendly}(\text{Bill}, \text{Susan}))$
3. $\text{True}(1, \text{Know}(\text{Bill}, \text{True}(1, \exists a (\text{Know}(\text{Susan}, \text{True}(1, \text{Friendly}(a, \text{Susan}) \wedge \text{Know}(a, \text{comb}(\text{Sf1}, 1))))))))))$

Bill cannot plan to ask the third agent for the combination of the safe, since the third agent might be unfriendly and refuse to cooperate. Rather, he can plan to ask Susan for the combination. This strategy should work: Susan should be able to learn the combination, presumably by asking the third agent, and should then be able to give over the information to Bill. Bill's plan thus looks something like this:

- Bill ask Susan for the combination
- (Susan learns the combination, possibly as follows:
 - Susan asks the third agent for the combination
 - the third agent tells Susan the combination

- Susan tells Bill the combination

This type of plan is known as a *vicarious* plan [Morgenstern, 1988]. Bill, who is constructing the plan, does not prepare all the details of his plan. Rather, he delegates a large part of his plan to Susan. He tells Susan some general constraints on this part of the plan: i.e., he tells Susan that she should learn the combination of the safe, but he does not tell her exactly how she should do it. In fact, he does not know exactly how she will learn the combination. Implicit in his delegating a high-level task to Susan is his inevitable loss of knowledge of and control over the fine points of the plan. Traditional planners, such as STRIPS [Fikes and Nilsson, 1971] and TWEAK [Chapman, 1985] cannot express vicarious plans. Thus, they are not expressive enough to describe the sort of plan needed in Example 3.

There are two important points to note here. The first is that we once again run into the Third Agent Frame Problem. Bill cannot prove that Susan's plan will work: that the third agent will still know the current combination of the safe when Susan asks him (because the third agent will not necessarily know that Bill has asked Susan to get the combination). Let us assume for the moment, however, that we are somehow able to get around this problem; that Bill is able to prove that Susan's plan will succeed, and thus can prove that his own vicarious plan will succeed. Even if Bill can prove that his plan to find out the combination of the safe will work, he will still not be able to prove that he can dial the combination of the safe in the resulting situation, because he does not know that he will still be near the safe in the resulting situation. This is because Bill does not know exactly what goes on between the time that he asks Susan for the information and the time Susan gives him the information. He can surmise what the plan might be; he can prove that it is *possible* for Susan to get the information, and that therefore she does get the information (Axiom Goals3), he can even suggest a possible plan that Susan might use (asking the third agent for the combination, and the third agent telling Susan the combination), but he cannot know exactly what goes on. It is possible that Susan might execute some very strange plan in order to obtain the combination which would result in his no longer being near the safe. For example, Susan might ask Bill to leave the room while she asks the third agent for the combination, or she might decide to learn the combination by taking apart the safe.

Therefore, he cannot apply the frame axioms that would let him conclude that he is still at the safe, and that he is thus still physically capable of opening the safe.

We will refer to this problem as the Vicarious Planning Frame Problem. In its most general form, it is the problem of an agent determining what stays the same about the world when he delegates part of his plan to other agents. In realistic planning situations, where plans are often complex, and agents typically must delegate parts of their plan, must relinquish control to other agents, and cannot stay on top of every detail, this problem is widespread. Once again, this is a problem that cannot be solved by throwing more frame axioms into the theory. Once again, it is a problem that arises even in a theory that does not allow for concurrent actions.

In the next section, we discuss the underlying reasons behind these two variant frame problems, and show why standard solutions to the original temporal projection problem will not work here.

4 Why Standard Solutions Won't Work

4.1 The Problem With Standard Temporal Logics

Perhaps the most salient feature of both the Third Agent and Vicarious Planning Frame Problems is that they are totally immune to an attack by frame axioms. What is missing in both cases is not frame axioms, but some agent's ability to apply these frame axioms. In both cases, the agent cannot apply the frame axioms because he does not know all the actions that have taken place during some interval of time.

Actually, the fact that frame axioms will not solve these variant frame problems can be seen as an artifact of the particular way in which the original solution to the temporal projection problem was set up. McCarthy and Hayes, when they suggested frame axioms as a solution to the temporal projection problem, considered single-agent systems which did not allow for concurrency: i.e., systems in which one agent did one action at a time.

⁶ Moreover, in these systems, one always began with some initial time point, or situation,

⁶This is implicit, though not explicit, in the paper. Later versions of the situation calculus took this implicit assumption to be an immutable property of the situation calculus.

and described later time points by specifying the complete history of the world since the initial situation. Such systems would reason about which facts stayed the same during a period of time which satisfied the following two properties:

- (1) For any two time points contained in the time period, it was a theorem of the system that at least one action or subaction took place during that time (i.e. the system always knew of something that occurred at any particular time: there were no gaps in the system's knowledge)
- (2) If an action occurred during the time period, it was a theorem of the system that that action occurred (i.e. all actions were known to the system.)

If a time period satisfies the first property, we say that it is *dense*; if it satisfies the second property, we say that it is *complete*. We define a system to be dense if all its time periods are dense; it is complete if all its time periods are complete. It can easily be seen that all time periods of any instance of the standard situation calculus (i.e., the time between any two situations) are dense and complete.

In a dense and complete system, frame axioms are clearly adequate to handle the temporal projection problem. In particular, since the system knows of all actions that take place, and since there is never a period of time in which the system does not know of something that is happening, the system always “knows enough” to apply the frame axioms in any situation.

We now expand the definitions of density and completeness to multiple agent systems. We say that a time period T of a system is dense with respect to a particular agent A if for any time period S , $S \subset T$, A knows of some action or subaction that occurs during S . T is complete with respect to A if, whenever an action occurs during T , A knows that the action occurs. A system is dense with respect to a particular agent if all time periods of the system are dense with respect to that agent; a system is complete with respect to a particular agent if all time periods of the system are complete with respect to that agent.

If a system is dense and complete with respect to a particular agent, and if the agent knows all relevant frame axioms, he will be able to handle the temporal projection problem.

In particular, he will always know enough to apply the frame axioms in any situation. Consider, however, Examples 2 and 3, above. In Example 2, the interval between times 1 and 3 is neither dense nor complete from the point of view of the third agent. In Example 3, the interval between time 2 and time 4 is neither dense nor complete from Bill's standpoint, and the interval between time 1 and time 2 is neither dense nor complete from the third agent's viewpoint. Thus they do not have sufficient information to apply these frame axioms.

In fact, however, any solution to the frame problem should be independent of the properties of density and completeness. After all, neither density nor completeness should play any role in one's ability to perform temporal projections. Consider, for example, my reasoning in the following situation:

I have just bought a ready-to-assemble desk at IKEA, and I plan to assemble the desk in my study. I open the package, take out all the parts, screws, and nails, and place them on the floor of my study. I realize that the package does not contain an Allen wrench, which I will need to assemble some of the desk parts. I quickly run down to the hardware store to buy an Allen wrench. As I reenter my home, I expect that the desk parts, nails, and screws are still on the floor. I come to this conclusion not because I know of all the events that have transpired since I placed the parts on the floor, and certainly not (even if I do know of all the events) because I am reasoning about all the events that have taken place. It seems more likely that I conclude that the parts are on the floor because I do not know of any action that has taken place which would cause the parts to move. That is, the removal of the desk parts is *unmotivated*.

Similarly, consider the fact that we all go to sleep at night and wake up each morning, largely unaware of everything that has happened during the night. Nevertheless, we assume that we will still find the cereal in the kitchen cabinet and the milk in the refrigerator and that the car will still be in the garage. Again, the changes in location of the cereal, milk, or car are unmotivated. We will return to this point in the next section. The main point of the argument, at present, is that density and completeness are properties which are superfluous to the temporal projection problem.

4.2 The Problem With Non-monotonic Temporal Logics

At first glance, it might seem that the Third Agent and Vicarious Planning Frame Problems arise only within monotonic logics. Clearly, it is unrealistic to expect that the third agent will be able to know with absolute certainty that the combination of the safe has not changed, or that Bill will predict that he will definitely still be near the safe when he learns the combination. Rather, these are plausible inferences that Bill and the third agent will make. It would seem, then, that a non-monotonic temporal logic would more accurately model the sort of reasoning used by agents in multiple-agent planning systems.

Indeed, many researchers ([McDermott, 1982], [McCarthy, 1986]) have suggested that the proper approach to the temporal projection problem — and one that would obviate the need for frame axioms — would be based on some sort of default or non-monotonic reasoning. McDermott has noted that properties typically persist for varying durations of time and has suggested that one associates with each property a persistence of a certain length. For example the persistence of the property of a cat being on the sofa might be measured in minutes; the persistence of a boulder being on a road might be measured in years. In theory, this approach seems promising: for Example 2, one could associate a persistence of years to the fluent $\text{comb}(\text{Sf1})$ (the combination of the safe), assume that the communicative actions take relatively short periods of time, and that all agents know these facts. This might permit the conclusion that the third agent knows the combination after Bill asks him for it. Unfortunately, McDermott's approach has not been developed in sufficient detail so that we can make concrete claims that it solves the variant frame problems.⁷

McCarthy [1986] has presented a detailed solution to the temporal projection problem within a non-monotonic logic. His solution is developed within the situation calculus. The basic idea behind his approach is as follows: One wishes to assume that properties persist for as long as possible. To achieve this result, one assumes that an action typically does not change the value of a fluent (i.e., a property that changes over time) unless it is *abnormal* in a particular respect. For example, consider the action $\text{Move}(a,b)$. This action would be

⁷Recently, Haugh [1989] has begun work on a theory of temporal reasoning which makes use of McDermott's theory of persistences. This work is still preliminary.

abnormal with respect to the fluent `location(a)` but would not be abnormal with respect to the fluent `color(a)`. One then circumscribes the abnormality predicate, so that its extension is as small as possible. Thus, unless one can deduce that a particular action is abnormal to some fact, one non-monotonically concludes that it is not abnormal, and that indeed, the fact remains true after the action has been performed.

McCarthy's original formulation fell victim to the *multiple extension problem*; that is, a set of axioms in McCarthy's circumscriptive logic can give rise to multiple models, or extensions. This can be problematic when some of these models appear to contradict our commonsense understanding of the axioms. A special case of this is the Yale Shooting Problem [Hanks and McDermott, 1986], which was discovered when Hanks and McDermott attempted to integrate temporal and non-monotonic logics. The Yale Shooting Problem can be briefly described as follows: Assume a simple model of time, where actions always take unit time. We are told that a gun is loaded at time 1, and that the gun is fired at Fred at time 5. Loading a gun causes the gun to be loaded, and firing a loaded gun at an individual causes the person to be dead. In addition, the fluents "alive" and "loaded" persist as long as possible; i.e., a person who is alive tends to remain alive, and a gun that is loaded tends to remain loaded. What can we conclude about Fred's status at time 6? If we are working within the situation calculus,⁸ and we assume that the starting situation is `S0`, we phrase the question: Is

```

Holds(Alive (Fred),
      Result (Shoot,
             Result(Wait,Result(Wait,Result(Wait,Result(Load,S0))))))
true?

```

Although common sense argues that Fred is dead at time 6, the facts support two mod-

⁸The Yale Shooting Problem was originally formulated within the situation calculus. This fact has greatly affected the proposed solutions to the YSP, as discussed below. Nevertheless, Hanks and McDermott originally viewed the YSP as a problem that arises in any temporal formalism [personal communication]; it is not unique to the situation calculus. We thus feel free to refer to the recasting of the Yale Shooting Problem in a generic temporal formalism as the Yale Shooting Problem.

els. In one model (the expected model), the fluent “loaded” persists as long as possible. Therefore, the gun remains loaded until it is fired at Fred, and Fred dies. In the other, unexpected model, the fluent “alive” persists as long as possible; i.e., Fred is alive after the shooting. Therefore, the fluent “loaded” did not persist; somehow the gun must have become unloaded.

Any solution to the Yale Shooting Problem must somehow disqualify these unexpected models. Modifications to the McCarthy system which solve the YSP include those developed by Lifschitz [1986,1987], Haugh [1987], Kautz [1986], and Baker and Ginsberg [1989]. Some of these solutions ([Lifschitz, 1986], [Kautz, 1986]) work by imposing a forward-in-time order on reasoning. In the expected model, one reasons from the earliest to the latest time point; in the unexpected model, one reasons from the latest time point (after the shooting) to earlier times. Thus, the unexpected models are disqualified and we can conclude that Fred is dead after the shooting. These solutions for the most part have fallen out of favor because they do not support the backward reasoning that is so important for belief revision and explanation. (See Section 4.3 for more on this point.) Other solutions to the Yale Shooting Problem ([Lifschitz, 1987], [Haugh, 1987]) work by introducing a predicate `cause` and circumscribing this predicate: this ensures that an action will not cause any unexpected effects to appear. For example, circumscribing `cause` means that the `Wait` action in the YSP will not cause the gun to become unloaded.

These solutions successfully handle the Yale Shooting Problem (although they do not necessarily work for related problems); nevertheless, they, along with McCarthy’s original non-monotonic temporal logic, cannot serve as the basis for the solution to the Third Agent and Vicarious Planning Frame Problems. This is primarily because all of the solutions cited above are based on the situation calculus. Standard situation calculus is in fact not even expressive enough to express the Third Agent and Vicarious Planning Frame Problems, as mentioned in Section 3.3. The situation calculus was originally formulated to represent single-agent plans where every plan step and every action are completely specified. A representation for multiple agents as well as for partially specified actions is needed for both variant frame problems. In addition, a representation for partial plans is needed for the Vicarious Planning Frame Problem.

One might argue that this objection is solvable; that one would merely have to extend the situation calculus so that it is more expressive. Some efforts in this direction have already been made by [Gelfond, Lifschitz, and Rabinov, 1991]. They extend the situation calculus so that it can express, among other things, concurrency and gaps, although it does not (yet) permit expression of the multiple agent plans discussed in section 3.3.

But the lack of expressivity of the situation calculus is not the only objection we have to those solutions to the Yale Shooting Problem that are based on this ontology. The more serious objection is that all of the solutions that are based on the situation calculus utilize its restrictions and inexpressivity in an essential manner. All of these non-monotonic temporal logics are dense, and with the exception of Haugh’s system⁹, they are all complete. One must know what actions have taken place at each situation in order to apply the basic principle of inertia underlying these systems and to reason that facts remain the same. Non-monotonicity is used, not to infer that “troublesome” actions (i.e. those that cause facts to change) have not occurred, but instead, to infer that the actions that have been described have not caused troublesome effects. Such inferences are necessary, but not sufficient, for a theory of temporal reasoning. Indeed, it is easy to see that Lifschitz’s [1987] solution to the Yale Shooting Problem fails for the extended situation calculus that is presented in [Gelfond, Lifschitz, and Rabinov, 1991].¹⁰ The problem is that Lifschitz, in his original solution to the Yale Shooting Problem, *used* the restrictive assumptions of the situation calculus in formulating his solution to the YSP. When these limiting assumptions are lifted,

⁹the first theory described in [Haugh, 1987]

¹⁰Consider, for example the following causal theory:

`Causes(load,loaded,true); Causes(unload,loaded, false); Causes(shoot, loaded, false); Causes(shoot,alive, false),: Precond(loaded,shoot), as well as the general axiom on affects, and the general principle of causation described in [Lifschitz, 1987]. The initial situation is described as holds(alive,S0); ¬ holds(loaded,S0). Now consider the following two axioms:`

`Axiom A: ∃ a S1 = result(shoot,result(a,result(load,S0)))`

`Axiom B: ∃ b S2 = result(shoot,result(wait + b, result(load,S0))`

where the + operator represents concurrency. Axiom A describes the YSP with gaps; Axiom B describes the YSP with concurrent actions. Lifschitz’s solution to the YSP no longer holds. There is no way to rule out models where a (or b) = unload, the gun is unloaded, and Fred lives. Therefore, we cannot predict that Fred is dead in S1 or S2.

his solution no longer holds.

Because of the critical misconceptions underlying these solutions to the Yale Shooting Problem, these non-monotonic temporal logics cannot be used for general temporal projection. Thus, they are not adequate for solving the Third Agent and Vicarious Planning Frame Problems. What is needed is a non-monotonic temporal logic based on sound principles that are independent of the peculiarities of a particular ontology, and that hold for temporal reasoning in general. Such a system follows below.

4.3 What Could Work: Proposal for a Non-Monotonic Temporal Logic

In the Allen wrench example above, it was argued that I reason that the desk parts and screws are still on the floor of the study when I return home because I do not know of any action that has taken place which would cause the parts to move from the floor. This is clearly a defeasible conclusion: I may return home and discover that the wind has caused one of the screws to roll outside the room, or that the housekeeper has placed the parts on a table in the room. Nevertheless, this seems to be a conclusion that is warranted by a rational system. Since there is no reason to believe that the wind or housekeeper has changed the location of the parts, we would like to develop a system that likewise will assume that unexpected actions do not happen. That is, we aim to develop a non-monotonic temporal logic that allows an agent, if he does not know of a particular action, to conclude that the action has not taken place.

As was mentioned above, most of the non-monotonic temporal logics that have so far been developed are not candidates for our purpose, due to their density and/or completeness. To our knowledge, there are only three reasonably developed systems that are candidates: Shoham's [1987] logic of chronological ignorance, Morgenstern and Stein's [1988,1989] Motivated Action Theory, and Amsterdam's [1991] theory of temporal reasoning. These theories all operate on the desired assumption: that as little unexpected happens as is possible. In Shoham's theory, as in the theories of [Lifschitz, 1986] and [Kautz, 1986], this is expressed by requiring changes to happen as late as possible. In Morgenstern's and Stein's theory as well as in Amsterdam's theory, this is expressed by requiring that we prefer models in which actions happen only if they are *motivated* - i.e., if they *have* to happen, given the agent's

current knowledge of what has happened in the world. These systems are neither dense nor complete. They are not based on the situation calculus, and it is possible to express an agent's incomplete account of the events in some chronicle. Shoham's theory, however, is flawed in other respects. In particular, like the systems described in [Lifschitz, 1986] and [Kautz, 1986], it forces reasoning to go forward in time. Thus, it is unsuitable for a theory of explanation. It gives strange and unexpected results for problems very similar to the Yale Shooting Problem. For example, given the original Yale Shooting Problem, together with the fact that Fred is *alive* at time 6, Shoham's theory would conclude that someone had unloaded the gun at time 4, the last possible moment. Common sense, on the other hand, concludes that someone must have unloaded the gun, but cannot conclude whether the gun became unloaded at time 2, time 3, or time 4. In general, Shoham's theory would be incapable of performing backward projection in any reasonable manner. In contrast, the theories of Morgenstern and Stein and of Amsterdam are good candidates which support both forward and backward reasoning. Amsterdam's theory, however, is not as fully developed as Motivated Action Theory; in addition, his work postdates most of the research done in this paper.

In the next section, we describe Morgenstern and Stein's logic and extend it so that it can handle the Third Agent Frame Problem and the Vicarious Planning Frame Problem.

5 Extending Motivated Action Theory

5.1 Synopsis of Motivated Action Theory

Motivated Action Theory (MAT) is based on the principle that an agent typically knows all that he needs to know in order to make predictions about the world in which he lives. In particular, if he is reasoning with an underconstrained set of facts about the world, he will tend to prefer models which have fewer unexpected, *unmotivated* actions. For example consider again the Yale Shooting Problem, presented in the previous section. Suppose that an agent knows that someone loads a gun at time 1 and shoots the gun at Fred at time 5, and he also knows that shooting a loaded gun at someone causes that person to die. The agent should predict that Fred is dead at time 6. He does not typically consider the

possibility that someone unloads the gun between times 2 and 5, and that Fred is therefore alive, because the unload action would be *unmotivated* in this context.

MAT¹¹ assumes a simple first order logic \mathcal{L} . For simplicity, we let the syntax of \mathcal{L} be the same as the syntax of \mathcal{E} . A theory T and a chronicle description CD are sets of sentences of \mathcal{L} . Intuitively, T gives the rules governing the world's behavior; CD describes some of the facts that are true and the actions that occur during a particular interval of time. $T \cup CD = TI$, a *theory instantiation*. A theory T contains causal rules and persistence rules. Causal rules describe how actions change the world; persistence rules describe how fluents remain the same over time. Causal rules are of the form

$$\alpha \wedge \beta \Rightarrow \gamma$$

where α is the set of triggering events of the causal rule, β gives the preconditions of the action, and γ describes the results. Persistence rules are of the form

$$\text{True}(t, f) \wedge \beta \Rightarrow \text{True}(t+1, f)$$

where β does not include statements of the form $\text{True}(t, \text{occurs}(\text{act}))$. Persistence rules are instances of the principle of inertia, (any fluent that holds at a time point t holds at $t+1$, unless some action which causes the fluent to change its truth value occurs at time t). In general, they do not have to be hand coded for each theory, although we will be hand coding the persistence rules needed for the simple examples in this paper. A chronicle description CD contains the boundary conditions and occurrences of a particular theory instantiation.

Intuitively, an action is motivated with respect to a theory instantiation if there is a “reason” for it to happen.¹² An action is strongly motivated if it “has to happen” in all models, i.e., it is a theorem that the action happens. It is weakly motivated if it “has to happen” with respect to a particular model, i.e., if it must occur given the particular way

¹¹The original version of MAT, given in [Morgenstern and Stein, 1988] was revised in [Stein and Morgenstern, 1989]. Several problems with both of these versions came to light during the current research. The version given here differs in a few respects from the previous versions. These differences are pointed out in the text.

¹²More precisely, it is statements, not actions, that are motivated. Informally, we say that an action act is motivated at time t if the statement $\text{True}(t, \text{Occurs}(\text{act}))$ is motivated.

the model is set up. It is semi-motivated if we are explicitly told that either the action happened, or another fact is true. (Note, in this case, that just because an action is semi-motivated does not necessarily mean that it occurs.) It is existentially motivated if it is a skolemized instance of an existentially quantified statement that is motivated.

Definition 2

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

(b) A statement ϕ is strongly motivated with respect to TI if ϕ is in all models of TI . If ϕ is strongly motivated with respect to TI , we say that it is motivated in $\mathcal{M}(TI)$, for all $\mathcal{M}(TI)$.

(c) A statement ϕ is weakly motivated with respect to $\mathcal{M}(TI)$ if there exists in TI a ($[n]$ instance of a) causal or persistence rule of the form $\alpha \wedge \beta \Rightarrow \phi$, α is (strongly or weakly or semi or existentially) motivated in TI and $\mathcal{M}(TI) \models \beta$.

(d) A statement ϕ is semi-motivated in TI if ϕ is of the form $0\text{occurs}(t, \text{act})$ and $\phi \vee \psi_1 \vee \dots \vee \psi_n \in CD$. If ϕ is semi-motivated with respect to TI , we say that it is semi-motivated in $\mathcal{M}(TI)$, for all $\mathcal{M}(TI)$.

(e) A statement ϕ is existentially motivated with respect to $\mathcal{M}(TI)$ if there exists some statement ψ such that ψ is motivated in $\mathcal{M}(TI)$, and ϕ is obtained from ψ in the following manner: ψ is of the form $\exists x \rho(x)$ and ϕ is $\rho(S)$ where S is some skolem constant.

It is an immediate consequence of strong motivation that if there is a statement of the form $\forall x \text{True}(t, 0\text{occur}(f(x)))$ then, for all constants A , $\text{True}(t, 0\text{occurs}(f(A)))$ is strongly motivated.

The concept of existential motivation is new to this version of MAT. Quantification was ignored in the original version of MAT [Morgenstern and Stein, 1988]. In an intermediate version of MAT [Stein and Morgenstern, 1989], if an existentially quantified statement - e.g. $\exists x \text{True}(t, 0\text{occurs}(f(x)))$

was present in the CD, it was considered as if the infinite disjunction

$f(a1) \vee f(a2) \vee \dots \vee f(a_n) \dots$

existed in the CD. Either approach could lead to anomalous results. In particular, consider the following theory containing a causal chain:

CD:

$\exists x \text{ True}(1, \text{Occurs}(f(x)))$

T:

$\text{True}(t, \text{Occurs}(f(x))) \wedge \text{True}(t, s) \Rightarrow \text{True}(t+1, \text{Occurs}(g(x)))$

One would expect the theory to project

$(\exists x \text{ True}(2, \text{Occurs}(g(x)))) \Leftrightarrow \text{True}(1, s)$.

However, the original version of MAT would project $\neg \text{True}(1, s)$ *unless* one allows there to be unnamed objects in the model (i.e. we assume that the mapping between the language and the domain is not onto). For, for any a , $\text{True}(1, \text{Occurs}(f(a)))$ is unmotivated and $\text{True}(2, \text{Occurs}(g(a)))$ is unmotivated. Since MAT minimizes unmotivated actions, it would prefer models in which $\neg \text{True}(1, s)$ and thus g does not occur at time 2.¹³

The second version of MAT gives even more anomalous results. Consider the following problem, which we will call the Nightmare Frame Problem.

CD:

$\text{True}(1, f)$ [where f is any fluent]

$\exists \text{ act True}(1, \text{Occurs}(\text{act}))$

T:

$\text{True}(t, \text{Occurs}(\text{act}_1)) \Rightarrow \text{True}(t+1, \text{Result}_1)$

...

$\text{True}(t, \text{Occurs}(\text{act}_n)) \Rightarrow \text{True}(t+1, \text{Result}_j)$

If one treats each instantiation of the existential statement as motivated, then if there is any action that results in f being made false, we will not be able to conclude anything about f at time 2. More generally, we lose all information about the world whenever some unspecified action is performed. Such behavior is clearly very problematic.

The version given here does not suffer from these anomalies and treats existential quan-

¹³This problem should not arise in MAT if we assume unnamed entities in the models of a theory; that is, we assume that the function mapping the language to the domain is not onto. If we assume this, the original version of MAT should be equivalent to what is presented in this paper.

tifiers in an intuitive manner.

A model is preferred if it has as few unmotivated actions as possible. To formalize this concept, we introduce the following functions on models:

- $Occurs\text{-}statements(M(TI))$, the set of statements ϕ , where ϕ is of the form $\text{True}(\tau, \text{occurs}(\text{act}))$ and $\mathcal{M}(TI) \models \phi$
- $Mot(M(TI))$, the set of statements in $Occurs\text{-}statements(M(TI))$ which are motivated in $\mathcal{M}(TI)$
- $Unmot(M(TI))$, the set of unmotivated statements in $\mathcal{M}(TI)$,
 $= Occurs\text{-}statements(M(TI)) - Mot(M(TI))$

We now have the following definitions:

Definition 3 $\mathcal{M}_\gamma(TI) < \mathcal{M}_\beta(TI)$ (\mathcal{M}_γ is preferable to \mathcal{M}_β) if $Unmot(\mathcal{M}_\gamma) \subseteq Unmot(\mathcal{M}_\beta)$

14

Definition 4 $\mathcal{M}_\gamma(TI)$ is as preferred as $\mathcal{M}_\beta(TI)$ ($\mathcal{M}_\gamma(TI) \approx \mathcal{M}_\beta(TI)$) if $\mathcal{M}_\gamma(TI) < \mathcal{M}_\beta(TI)$ and $\mathcal{M}_\beta(TI) < \mathcal{M}_\gamma(TI)$.

Definition 5 $\mathcal{M}(TI)$ is a preferred model for TI if

$$\mathcal{M}'(TI) \leq \mathcal{M}(TI) \Rightarrow \mathcal{M}'(TI) \approx \mathcal{M}(TI).$$

Definition 6 Let $\mathcal{M}^*(TI)$ be the set of all preferred models. Let $\cap_{\mathcal{M}^*(TI)} = \{\phi \mid \forall \mathcal{M} \in \mathcal{M}^*(TI), \mathcal{M} \models \phi\}$ - the set of statements true in all preferred models of TI.

Definition 7 If $\phi \in \cap_{\mathcal{M}^*(TI)}$, we say TI $\rightsquigarrow \phi$, to be read as: TI projects ϕ .

MAT can be used for both forward temporal projection — reasoning about what will be true in the future, and backward temporal projection — reasoning about what must have been true in the past. Thus, given the Yale Shooting Scenario, MAT will project that Fred is dead at time 5. If MAT is subsequently given the information that Fred is alive at

¹⁴This definition is different from the definition given in [Morgenstern and Stein, 1988] and [Stein and Morgenstern, 1990]. The preference criterion there was incorrect, and was not in fact transitive. Thanks to Ramiro Guerreiro and Jonathan Amsterdam for pointing this out, and for Ramiro Guerreiro for suggesting the correct preference criterion, above.

time 6, MAT will infer that an unload must have occurred at time 2, 3, or 4. In fact, it can be shown that MAT gives the *simplest possible explanation*, i.e., the explanation that it projects for an unexpected occurrence contains the fewest number of actions.

The deduction theorem holds in MAT. That is, if $TI \cup b \rightsquigarrow p$, then $TI \rightsquigarrow b \Rightarrow p$. This is a direction consequence of Shoham [1987], who shows that the deduction theorem always holds in model-preference theories. This is a very useful result, since it allows us to use natural deduction proofs within MAT.

5.2 Integrating MAT with an epistemic logic

The principle underlying MAT — that one prefers models in which the fewest unmotivated actions take place — seems central to the sort of reasoning that one needs to solve examples 2 and 3. Presumably, in Example 2, Bill reasons as follows: It is true that the third agent will not know what took place between times 1 and 3. But probably, nothing happened to change the safe’s combination during that time. Since nothing untoward happened, the third agent will not know that anything untoward happened, and will therefore reason (using the principle underlying MAT) that nothing untoward happened. Therefore, he will reason that he knows the combination, and everything will run smoothly.

In Example 3, Bill presumably reasons: I don’t know of anything Susan would do in her plan to learn the combination of the safe that would change my proximity to the safe. Therefore, I will assume that these conditions will hold true when I find out the combination of the safe.

In these examples, Bill reasons that he, the third agent, and Susan will prefer models in which unmotivated actions do not happen. This reasoning will allow Bill to successfully plan to perform an action.

While the basic principle underlying MAT seems to be just what is needed, MAT by itself cannot handle examples 2 and 3. To handle these examples, we will have to integrate MAT with an epistemic temporal logic such as \mathcal{E} which was presented in section 2. In particular, we will have to show how individual agents can reason using MAT. We will be especially interested in incorporating nested levels of reasoning within MAT. That is, we will have to show how an agent can reason that another agent is making predictions using

the principles underlying MAT. We present this integrated theory, called EMAT (Epistemic Motivated Action Theory), below.

5.2.1 Epistemic Motivated Action Theory

The language underlying EMAT is an instance of the first order predicate calculus. In many respects, EMAT is quite similar to \mathcal{E} , the epistemic temporal logic introduced in section 3.1. There are, however, several important differences. One major difference concerns our choice of epistemic predicate. Clearly, an agent's reasoning takes place within the scope of an epistemic operator. In an ideal world, with infallible reasoning mechanisms, Know is an appropriate choice for this epistemic predicate. Know is no longer appropriate, however, if agents are reasoning using MAT. This is because MAT is a non-monotonic logic; its projections are defeasible. Thus, even if an agent *knows* his assumptions, at times he can at most be said to *believe* his conclusions. We therefore choose to replace Know with Bel (Believe) as the epistemic operator of choice. A theory of planning based on belief is considerably weaker than one based on knowledge. In particular, if an agent is constructing plans based on his *beliefs* about the world, as opposed to his *knowledge* about the world, he cannot predict with certainty that his plans will succeed. This is, however, the price we have to pay for a theory that can handle non-trivial problems of temporal projection.

In order to formalize how agents reason with MAT, we introduce the concept of a theory instantiation relativized to an agent and instant of time. $TI(a,t)$, the theory instantiation TI relativized to agent a and time t , consists of all the propositions that a believes at time t .

Definition 8 $TI(a,t)$ is the set of sentences p where either

1. $\text{True}(t, \text{Bel}(a, p)) \in TI$
2. p is of the form $P(x/C)$, where C is a skolem constant, $P(x/C)$ denotes the formula obtained by substituting C for all instances of x , and $\exists x \text{ True}(t, \text{Bel}(a, P(x))) \in TI$

Let us consider some examples. Recall Example 2, restated in terms of EMAT (changing Know to Bel). We give the following definition for the predicate Bel-what-is, analogous with the definition for Know-what-is in Section 3.3:

Definition 9 $\exists p \text{ True}(t, \text{Bel}(a, \text{Standard-id}(p) \text{ and } \text{Equal}(p, \text{trm}))$
 $\Leftrightarrow \text{True}(t, \text{Bel-what-is}(a, \text{trm}))$.

We thus get the boundary conditions:

1. $\text{True}(1, \text{Friendly}(\text{Bill}, \text{Susan}))$
2. $\text{True}(1, \text{Bel}(\text{Bill},$
 $\quad \text{True}(1, \exists a \text{ Bel}(\text{Susan},$
 $\quad \quad (\text{True}(1, \text{Bel-what-is}(a, \text{comb}(\text{Sf1}, 1))))))$

Thus, $TI(\text{Bill}, 1)$ contains the sentence

$$\text{True}(1, \exists a \text{ Bel}(\text{Susan}, (\text{True}(1, \text{Bel-what-is}(a, \text{comb}(\text{Sf1}, 1))))))$$

If TI contained the sentence

$$\text{True}(1, \exists x \text{ Bel}(\text{Bill}, \text{Murderer}(x)))$$

$TI(\text{Bill}, 1)$ could contain the sentence

$$\text{Murderer}(C)$$

where C is a skolem constant.

Notice that we allow skolemization only if p is of the following form:

$$\text{True}(t, [Q \text{ arg1}] \dots [Q \text{ argn}] \text{ Bel}(a, P(\text{arg1}, \dots, \text{argn})))$$

where Q stands for either \forall or \exists .

We do not allow skolemization if p is of the form

$$\text{True}(t, [Q \text{ arg1}] \dots [Q \text{ argn}]$$

$$R(\text{arg1}, \dots, \text{argn}) \text{ and } \text{Bel}(a, P(\text{arg1}, \dots, \text{argn})))$$

Using the concept of a relativized theory instantiation, we can define the concept of a nested relativization. $TI(a, t1, b, t2)$ refers to the beliefs that b has at time $t2$ from a 's point of view at time $t1$, relative to the theory instantiation TI . We thus have:

Definition 10 $TI(a, t1, b, t2) = TI(a, t1)(b, t2)$.

Thus, in our restatement of Example 2, above,

$TI(Bill,1,Susan,1) = TI(Bill,1)(Susan,1)$ could contain the statement

$True(1,Bel\text{-}what\text{-}is(A,Comb(Sf1,1)))$

where A is a skolem constant.

If TI contained the statement

$True(1,Bel(Elle n,True(2,Bel(Sandy,True(2,On(X,Y))))))$, then

$TI(Elle n,1,Sandy,2)$ contains the statement $True(2,On(X,Y))$.

Note from the above example that agent A can have a belief at time t_1 about the belief that another agent B has at a *later* time t_2 . We will often speak about what A at t_1 believes that B at t_2 *would believe* relative to some possible course of events. For example Susan at time 1 can reason about what she would believe at time 2 if she learns between time 1 and time 2 that she won the lottery (e.g., she might then believe that she could afford a Porsche, that she could afford to quit her job, etc.). To formalize this notion, we introduce the concept of the *extension of a theory instantiation*.

Definition 11 Consider a theory instantiation $TI1 = T \cup CD1$. Let $time\text{-}points(CD2)$ refer to the set of all time points mentioned in $CD2$. Let $TI2 = T \cup CD1 \cup CD2$, where $CD2$ has the following property:

If t_i is the latest time point in $CD1$, then $\forall t_j, t_j \in time\text{-}points(CD2) \Rightarrow t_j > t_i$

Then $TI2$ is an extension for $TI1$.

If $TI2$ is an extension of $TI1$, we write $TI1(a,t)_{[TI2]}$ to refer to what agent a believes at time t relative to what has occurred during $TI2$.

In order to allow agents to reason with MAT, we must define the concepts of motivation and projection with respect to a particular agent and time. A statement is motivated with respect to a particular agent and time if, from the point of view of the agent at that time, the statement has to be true. The formal definition is analogous to the definition of motivation discussed in Section 4.1.

Definition 12 (a) Let $TI(a,t)$ be a theory instantiation with respect to a and t . A statement ϕ is motivated in $\mathcal{M}(TI(a,t))$ if it is strongly motivated in $\mathcal{M}(TI(a,t))$, weakly motivated in $\mathcal{M}(TI(a,t))$, semimotivated in $\mathcal{M}(TI(a,t))$, or existentially motivated in $\mathcal{M}(TI(a,t))$.

(b) A statement ϕ is strongly motivated with respect to $TI(a,t)$ if ϕ is in all models of TI . If ϕ is strongly motivated with respect to $TI(a,t)$, we say that it is motivated in $\mathcal{M}(TI(a,t))$ for all $\mathcal{M}(TI(a,t))$.

(c) A statement ϕ is weakly motivated in $\mathcal{M}(TI(a,t))$ if there exists in $TI(a,t)$ a causal or persistence rule of the form α and $\beta \Rightarrow \phi$, α is (strongly, weakly, semi, or existentially) motivated in $\mathcal{M}(TI(a,t))$ and $\mathcal{M}(TI(a,t)) \models \beta$.

(d) A statement ϕ is semi motivated in $\mathcal{M}(TI(a,t))$ if ϕ is of the form $\text{True}(t, \text{occurs}(\text{act}))$ and $\phi \vee \psi_1 \vee \dots \vee \psi_n \in TI(a,t)$.

(e) A statement ϕ is existentially motivated in $\mathcal{M}(TI(a,t))$ if there is some statement ψ such that ψ is motivated in $\mathcal{M}(TI(a,t))$, and ϕ is obtained from ψ in the following manner: ψ is of the form $\exists x \rho(x)$ and ϕ is $\rho(S)$, where S is some skolem constant.

The functions *Occurs-statements*, *Mot*, and *Unmot* are identical to those in MAT. Likewise, the preference criterion on models in EMAT is identical to the preference criterion in MAT. We define $\mathcal{M}^*(TI(a,t))$ to be the union of all preferred models of $TI(a,t)$. Analogously with MAT, we define the following sets:

$\cap_{\mathcal{M}^*(TI(a,t))} = \{\phi \mid \forall M \in \mathcal{M}^*(TI(a,t))[M \models \phi]\}$ - the set of statements true in all preferred models of $TI(a,t)$.

$\cup_{\mathcal{M}^*(TI(a,t))} = \{\phi \mid \exists M \in \mathcal{M}^*(TI(a,t))[M \models \phi]\}$ - the set of statements true in at least one preferred model of $TI(a,t)$.

In general, if a statement $\phi \in \cap_{\mathcal{M}^*(TI(a,t))}$, we say that $TI(a,t)$ projects ϕ , $TI(a,t) \rightsquigarrow \phi$. That is, at time t , a will predict ϕ .

The concept of an extension for a theory instantiation motivates the following axiom of EMAT:

Axiom 1 $TI1(a,t)_{[TI2]} \rightsquigarrow \phi \Leftrightarrow TI2(a,t) \rightsquigarrow \phi$

Because the deduction theorem holds in MAT and EMAT, we have the following metatheorem of EMAT:

Theorem 2 Assume $TI2(a,t) = TI1(a,t) \cup CD2$. Then $TI2(a,t) \rightsquigarrow \phi$ iff $TI(a,t) \rightsquigarrow CD2 \supset \phi$

Often an agent a at time t_1 will wish to reason about what a group of agents all believe at some time t_2 . That is, he wishes to reason about the intersection of the facts that he believes each of the members of the group will believe at time t_2 . To formalize this concept, we have the following definitions:

Definition 13 *Let G be a group of agents. Then,*

$$\cap M * (TI(a, t_1, G, t_2) = \{ \phi \mid \forall g \in G \forall M \in M * (TI(a, t_1, g, t_2), M \models \phi) \}$$

Definition 14 $\cup P(M * (TI(a, t_1, G, t_2) = \{ \phi \mid \exists g \in G \forall M \in M * (TI(a, t_1, g, t_2)) M \models \phi \}$.

The theory as it has thus far been developed allows agents to reason about how other agents perform temporal reasoning. For example, given the assumptions of Example 2, it is a theorem that Bill believes at time 1 that if between time 1 and time 4, Bill asks Susan for the name of the third agent, Susan tells him the name of the third agent, and he asks the third agent for the combination of the safe, then at time 4, the third agent will believe that the combination of the safe has a certain value. Moreover, it will then be a theorem that the third agent will tell him what he believes to be the combination, and that at time 5, Bill will believe that the third agent believes that the combination has some specific value X . In order for this information to be useful to Bill, however, Bill must somehow come to believe that the combination of the safe is X . How can Bill come to believe this?

This is a non-trivial problem. Interestingly, it does not arise in integrated theories of *knowledge* and action. If an agent A *knows* that an agent B *knows* some sentence P , then it is certainly the case that A knows P as well. For if B knows P , then P is true; so A will know that P is true. If belief is the epistemic operator, however, the situation is not analogous. If A *believes* that B *believes* some sentence P , it is not necessarily the case that A himself believes P . For A might think that B is in fact mistaken about the truth of P .

Clearly, this is not the situation in Example 2. Bill has a certain amount of trust in the third agent's belief about the value of the combination of the safe; that is why he is asking him in the first place. In general, it makes sense to say the following: Suppose an agent A trusts a group of agents G . Suppose further that one of these agents (B) believes P , and that A has no reason to disbelieve P . That is, he does not have $\neg P$ in his belief base; further, he does not believe that any of the agents in G believes $\neg P$. Then A will believe P .

This (defeasible) inference rule is formalized as follows:

We assume a three place predicate Trust , where $\text{True}(t, \text{Trust}(a, b))$ means that a trusts b at time t . We then have the following inference rule.:

Nested Projections Inference Rule

Let $G = \{g \mid \text{Trust}(a, g, t1)\}$.

Assume

- (a) $p \in \cap M * (TI(a, t1, G, t2))$,
- (b) $\neg(\neg p \in \cap P(M * (TI(a, t1, G, t2))))$
- (c) $\neg(\neg p \in \cap M * (TI(a, t1)))$
- (d) p does not contain a skolem constant.

Then, $p \in \cap M * (TI(a, t1))$.

This rule is a restricted form of the Principle of Mutual Trust, formalized in [Davis, 1990: Chapter 8].

It is important to note that the inference rule is not just saying that A believes that which A believes B believes if A trusts B . The point is that in order for A to believe that which he believes B believes there cannot be any contradictory evidence from the other agents that A trusts. (To avoid trivializing the benchmark problems in this paper, we will assume in further discussion that A trust all agents, thus making the problem as difficult as possible.)

5.3 Results Using EMAT

The axioms for EMAT are given in Appendix B. These are identical to the axioms of Appendix A except that 1) The predicates Know , Know-what-is and $\text{Know-how-to-perform}$ are represented as Bel , Bel-what-is and $\text{Bel-how-to-perform}$, with the obvious interpretations, and 2) frame axioms are replaced by persistence rules. EMAT can handle Example 2. We can show that Bill can reason that the third agent will still know the combination of the safe when he asks for the number. That is, he can reason that the third agent will predict that the combination at time 3 is the same as the combination at time 1, since any action that would cause the combination to change is unmotivated. Therefore, he can reason that the third agent will give him the combination number, and that he will subsequently

know how to open the safe.

Likewise, EMAT can handle Example 3. We can show that Bill can reason that Susan can execute a plan that will result in her knowing the combination of the safe: namely, the plan of asking the third agent for the number, and the third agent telling Susan the number. The third agent will still know the combination of the safe when Susan asks him because any change to the combination is unmotivated. Similarly, Bill can predict that he will still be near the safe when he learns the combination, because any action that would result in his not being near the safe is unmotivated.

Note that if EMAT were based on the intermediate version of MAT, it could not handle Example 3. In particular, it would run into problems that were very similar to the Nightmare Frame Problem, described in Section 5.1. Specifically, Bill would not be able to prove that in the final situation, he would still be near the safe. All he knows of Susan's actions is that she executes *some* plan to learn the combination. Thus, in any model, the plan she executes would have some justification and would be motivated.

That is, he knows the statement:

$$\exists \text{pln } (T(2, \text{Can-execute-plan}(\text{Susan}, \text{pln})) \wedge T(2, \text{Occurs}(\text{pln})) \wedge \dots)$$

If this statement is interpreted as the infinite disjunction

$$T(2, \text{Can-execute-plan}(\text{Susan}, \text{Pln1})) \wedge T(2, \text{Occurs}(\text{Pln1})) \wedge \dots \vee$$

$$T(2, \text{Can-execute-plan}(\text{Susan}, \text{Pln2})) \wedge T(2, \text{Occurs}(\text{Pln2})) \wedge \dots \vee$$

...

$$T(n, \text{Can-execute-plan}(\text{Susan}, \text{Plnn})) \wedge T(2, \text{Occurs}(\text{Plnn})) \wedge \dots \vee \dots$$

...

then, every plan that results in Susan learning the combination, no matter how strange, would be (semi) motivated. This would include anomalous plans that can change the location of the safe. Using the current version of EMAT, however, these anomalies do not occur. EMAT handles these problems successfully.

In summary:

We have demonstrated that integrating an epistemic logic with a reasonably expressive theory of action and planning results in difficult variants of the temporal projection problem. These problems cannot be handled using traditional strategies. In particular, neither

old fashioned frame axioms, nor standard non-monotonic temporal logics can solve these problems.

We have integrated a system of non-monotonic temporal reasoning called Motivated Action Theory with a simple theory of knowledge and planning, and have shown that this theory can solve both variant frame problems. Some of the central concepts of this theory include the relativization of a theory instantiation to an agent and an instant of time, and the ability of agents to reason using nested, multi-agent beliefs.

The research described in this paper has served as the springboard for a more general investigation into multiple agent non-monotonic logics. The examples discussed here represent the first effort, as far as the author knows, to examine how agents reason about how other agents reason non-monotonically. In [Morgenstern, 1990], we examine general multiple agent non-monotonic logics. We extend Moore's [1985] autoepistemic logic to the multiple agent case, and show that the resulting logic is too weak to handle most reasonable problems of commonsense reasoning. We then suggest several inference rules, which allow an agent to be *arrogant* with respect to another agent's ignorance. While these principles of arrogance are in general too strong, restricted versions work quite well for commonsense reasoning. In particular, we show that a restricted form of one of the principles of arrogance is equivalent to the principle underlying EMAT.

We plan to extend EMAT in several directions. Our current treatment of quantifying in is insufficiently expressive for all planning problems, since it does not allow general quantification into epistemic contexts. In particular, we cannot express problems in which agents plan to combine their partial knowledge in order to perform some action. Finding some method for handling quantification into epistemic contexts is thus an important research area. In addition, we would like to develop a proof theory for EMAT, since currently, proofs are done on the model theoretic level.

6 Acknowledgements:

I have benefited greatly from discussions with, and the suggestions of, Ernie Davis, Ramiro Guerreiro, Hector Geffner, Wlodek Zadrozny, Kurt Konolige, Bob Moore, Lynn Stein, and Kate Sanders.

References

- [Amsterdam, 1991] Amsterdam, Jonathan: “Temporal Reasoning and Narrative Convention,” *Proceedings, KR 1991*
- [Baker, 1989] Baker, Andrew: “A Simple Solution to the Yale Shooting Problem,” *Proceedings, KR 1989*
- [Baker and Ginsberg, 1989] Baker, Andrew and Matt Ginsberg: “Some Problems in Temporal Reasoning,” submitted to *Artificial Intelligence*, 1989
- [Gelfond, Lifschitz, and Rabinov: 1991] Gelfond, Michael, Vladimir Lifschitz, and Arkady Rabinov: “What are the Limitations of the Situation Calculus,” *Working Papers, AAAI 91 Symposium on Logical Formalizations of Commonsense Reasoning*
- [Haugh, 1987] Haugh, Brian: “Simple Causal Minimizations for Temporal Persistence and Projection,” *Proceedings AAAI 1987*
- [Kautz, 1986] Kautz, Henry: “The Logic of Persistence,” *Proceedings, AAAI 1986*
- [Konolige, 1982] Konolige, Kurt: “A First Order Formalization of Knowledge and Action for a Multi-Agent Planning System,” J.E. Hays and D. Michie, eds.: *Machine Intelligence 10*, 1982 Also SRI TR 232, 1980
- [Lifschitz, 1987] Lifschitz, Vladimir: “Formal Theories of Action,” *Proceedings, IJCAI 1987*
- [Lifschitz and Rabinov, 1989] Lifschitz, Vladimir and Arkady Rabinov: “Miracles in Formal Theories of Action,” *AIJ Research Note*, 1989
- [Lifschitz, 1986] Lifschitz, Vladimir: “Pointwise Circumscription: Preliminary Report,” *Proceedings AAAI 1986*
- [McCarthy, 1986] McCarthy, John: “Applications of Circumscription to Formalizing Common-Sense Knowledge,” *Artificial Intelligence*, 1986
- [McCarthy and Hayes, 1969] McCarthy, John and Pat Hayes: “Some Philosophical Problems from the Standpoint of Artificial Intelligence,” Bernard Meltzer, ed.: *Machine Intelligence 4*, 1969

- [McDermott, 1984] McDermott, Drew: "The Proper Ontology for Time," unpublished paper, 1984
- [McDermott, 1982] McDermott, Drew: "A Temporal Logic for Reasoning About Processes and Plans," *Cognitive Science*, 1982
- [Moore, 1980] Moore, Robert: *Reasoning About Knowledge and Action*, SRI TR 191, 1980
- [Morgenstern, 1988] Morgenstern, Leora: *Foundations of a Logic of Knowledge, Action, and Communication*, NYU Ph.D. Thesis, Computer Science Dept., 1988
- [Morgenstern, 1990] Morgenstern, Leora: A Formal Theory of Multiple Agent Non-Monotonic Logics, *Proceedings, AAAI 1990*.
- [Morgenstern and Stein, 1988] Morgenstern, Leora and Lynn Andrea Stein: "Why Things Go Wrong: A Formal Theory of Causal Reasoning," *Proceedings AAAI 1988*. Longer version printed as Stein and Morgenstern "Motivated Action Theory," Brown CS TR 89-12, 1989.
- [Shoham, 1986] Shoham, Yoav: *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence*, MIT Press, 1987

Appendix A

Below, the axioms of the simple epistemic temporal logic \mathcal{E} . All statements are assumed to be universally quantified unless otherwise noted. Recall from Section 3.1 that all quotation and quasi-quotation marks are suppressed.

Axiom Friends1:

$$\text{True}(t, \text{Friendly}(a, b)) \Leftrightarrow \text{True}(t, \text{Friendly}(b, a))$$

The friendliness relation is symmetric.

Axiom Friends2:

$$\text{True}(t, \text{Friendly}(a, b)) \Rightarrow \text{True}(t, \text{Know}(a, \text{True}(t, \text{Friendly}(a, b))))$$

If an agent is friendly with someone, he knows it.

Axiom Goals1:

$$\begin{aligned} &\text{True}(t, \text{Friendly}(a, b)) \wedge \text{Occurs}(t, \text{do}(a, \text{request}(b, \text{act}))) \\ &\Rightarrow \text{True}(t+1, \text{Has-goal}(b, \text{act})) \end{aligned}$$

If a and b are friendly and a asks b to do a certain action, b will subsequently have the goal of doing that action.

Axiom Goals2:

$$\begin{aligned} &\text{True}(t, \text{Has-goal}(b, \text{act})) \text{ and } \text{True}(t, \text{Can-perform}(b, \text{act})) \\ &\Rightarrow \text{Occurs}(t, \text{do}(b, \text{act})) \end{aligned}$$

If b has the goal of performing a certain action and can perform that action, he will perform that action.

Axiom Goals3:

$$\begin{aligned} &\text{True}(t, \text{Has-goal}(b, \text{act})) \wedge \\ &\exists \text{pln } (\text{length}(\text{pln}) = m \wedge \text{True}(t, \text{Can-execute-plan}(b, \text{pln})) \wedge \text{True}(t, \text{occurs}(\text{pln})) \\ &\Rightarrow \text{True}(t+m, \text{Can-perform}(b, \text{act}))) \\ &\Rightarrow \exists \text{pln}' (\text{length}(\text{pln}') = m' \wedge \text{True}(t, \text{Can-execute-plan}(b, \text{pln}')) \\ &\wedge \text{True}(t, \text{occurs}(\text{pln}')) \Rightarrow \text{True}(t+m', \text{Can-perform}(b, \text{act}))) \wedge \text{True}(t, \text{occurs}(\text{pln}')) \end{aligned}$$

If b has the goal of performing a certain action and can execute some plan which will result in his being able to perform that action, then b will execute some plan which will result in his being able to perform that action.

Axiom Execute1:

$$\text{True}(t, \text{Can-perform}(b, \text{act})) \Rightarrow \text{Can-execute-plan}(b, \text{do}(b, \text{act}))$$

An agent can execute the plan of his performing an action if he can perform that action.

Axiom Execute2:

$$\text{True}(t, \text{Know}(b, \text{True}(t, \text{occurs}(\text{pln})))) \Rightarrow \text{Can-execute-plan}(b, \text{pln})$$

An agent can execute a plan if he knows it will occur.

Axiom Execute3:

$$\text{True}(t, \text{Can-execute-plan}(b, \text{pln1}))$$

$$\wedge \text{True}(t, \text{Know}(b, \text{True}(t, \text{occurs}(\text{pln1})) \wedge \text{length}(\text{pln1}) = m) \Rightarrow$$

$$\text{True}(t+m, \text{Can-execute-plan}(b, \text{pln2}))) \Rightarrow$$

$$\text{True}(t, \text{Can-execute-plan}(b, \text{sequence}(\text{pln1}, \text{pln2})))$$

An agent can execute a sequence of two plans if he can execute the first and knows that as a result of executing the first, he will be able to execute the second.

Axiom Know-Perform1:

$$\text{True}(t, \text{Know-what-is}(a, \text{comb}(\text{safe}, t)))$$

$$\Rightarrow \text{True}(t, \text{Know-how-to-perform}(a, \text{dial}(\text{comb}(\text{safe}, t))))$$

An agent knows how to dial a safe if he knows what the combination of the safe is.

Axiom Can-perform1:

$$\text{True}(t, \text{Know-how-to-perform}(a, \text{dial}(\text{comb}(\text{safe}, t)))) \wedge \text{True}(t, \text{At}(a, \text{safe}))$$

$$\Rightarrow \text{True}(t, \text{Can-perform}(a, \text{dial}(\text{comb}(\text{safe}, t))))$$

An agent can perform the action of dialing a safe if he knows how to perform the action and is near the safe.

Axiom Can-perform2:

$$\text{True}(t, \text{Know-what-is}(a, p)) \wedge \text{True}(t, \text{Friendly}(a, b))$$

$$\Rightarrow \text{True}(t, \text{Can-perform}(a, \text{tell-what-is}(b, p)))$$

An agent can tell a friend what something is if he himself knows what it is.

Axiom Can-perform3:

$$\text{True}(t, \text{Can-perform}(a, \text{do}(a, \text{request}(b, \text{act}))))$$

Agents can always perform request acts.

Now, the frame axioms:

Axiom Frame1:

$$\text{True}(t, \text{Know}(a, p)) \wedge \text{True}(t, \text{Occurs}(ev)) \wedge \text{Communicative-event}(ev)$$

$$\Rightarrow \text{True}(t+1, \text{Know}(a, p))$$

Axiom Frame2:

$$\text{True}(t, \text{Know-what-is}(a, p)) \wedge \text{True}(t, \text{Occurs}(ev)) \wedge \text{Communicative-event}(ev)$$

$$\Rightarrow \text{True}(t+1, \text{Know-what-is}(a, p))$$

These axioms say that communicative actions do not change an agent's knowledge. Examples of communicative events are *request*, the action of requesting an agent to do something, and *tell-what-is*, the action of telling an agent what something is.

Axiom Frame3:

$$\text{True}(t, \text{Friendly}(a, b)) \wedge \text{True}(t, \text{Occurs}(ev)) \wedge \text{Communicative-event}(ev)$$

$$\Rightarrow \text{True}(t+1, \text{Friendly}(a, b))$$

Axiom Frame4:

$$\text{True}(t, \text{At}(a, loc)) \wedge \text{True}(t, \text{Occurs}(ev)) \wedge \text{Communicative-event}(ev)$$

$$\Rightarrow \text{True}(t+1, \text{At}(a, loc))$$

Communicative actions do not change the friendliness of agents or their locations . (Axiom Frame3 is flagrantly false for such actions as insulting. We ignore this issue here, since it is far beyond the scope of this paper.)

We also add the following principles of knowledge:

Veridicality:

$$\text{True}(t, \text{Know}(a, p)) \Rightarrow \text{True}(t, p)$$

If an agent knows something, it is true.

Positive Introspection:

$$\text{True}(t, \text{Know}(a, p)) \Rightarrow \text{True}(t, \text{Know}(a, \text{True}(t, \text{Know}(a, p))))$$

If an agent knows something, he knows that he knows it.

Consequential Closure:

$$\text{True}(t, \text{Know}(a, p)) \wedge \text{True}(t, \text{Know}(a, p \Rightarrow q)) \Rightarrow \text{Know}(a, q)$$

Agents know the logical consequences of their knowledge.

To make Example 1 work, we add necessitation on *all* the axioms, including frame axioms, and state that the hearer and listener of a communicative action always know when it has taken place.

Appendix B

Below, the axioms of Epistemic Motivated Action Theory \mathcal{EMAT} . All statements are assumed to be universally quantified unless otherwise noted. Recall from Section 3.1 that all quotation and quasi-quotation marks are suppressed.

Axiom Friends1:

$$\text{True}(t, \text{Friendly}(a, b)) \Leftrightarrow \text{True}(t, \text{Friendly}(b, a))$$

The friendliness relation is symmetric.

Axiom Friends2:

$$\text{True}(t, \text{Friendly}(a, b)) \Rightarrow \text{True}(t, \text{Bel}(a, \text{True}(t, \text{Friendly}(a, b))))$$

If an agent is friendly with someone, he believes that he is friendly with that person.

Axiom Goals1:

$$\text{True}(t, \text{Friendly}(a, b)) \wedge \text{Occurs}(t, \text{do}(a, \text{request}(b, \text{act})))$$

$$\Rightarrow \text{True}(t+1, \text{Has-goal}(b, \text{act}))$$

If a and b are friendly and a asks b to do a certain action, b will subsequently have the goal of doing that action.

Axiom Goals2:

$$\text{True}(t, \text{Has-goal}(b, \text{act})) \text{ and } \text{True}(t, \text{Can-perform}(b, \text{act})) \Rightarrow \text{Occurs}(t, \text{do}(b, \text{act}))$$

If b has the goal of performing a certain action and can perform that action, he will perform that action.

Axiom Goals3:

$$\text{True}(t, \text{Has-goal}(b, \text{act})) \wedge$$

$$\exists \text{pln } (\text{length}(\text{pln}) = m \wedge \text{True}(t, \text{Can-execute-plan}(b, \text{pln})) \wedge \text{True}(t, \text{occurs}(\text{pln}))$$

$$\Rightarrow \text{True}(t+m, \text{Can-perform}(b, \text{act}))$$

$$\Rightarrow \exists \text{pln}' (\text{length}(\text{pln}') = m' \wedge \text{True}(t, \text{Can-execute-plan}(b, \text{pln}'))$$

$$\wedge \text{True}(t, \text{occurs}(\text{pln}'))$$

$$\Rightarrow \text{True}(t+m', \text{Can-perform}(b, \text{act})) \wedge \text{True}(t, \text{occurs}(\text{pln}'))$$

If b has the goal of performing a certain action and can execute some plan which will result in his being able to perform that action, then b will execute some plan which will result in

his being able to perform that action.

Axiom Execute1:

$$\text{True}(t, \text{Can-perform}(b, \text{act})) \Rightarrow \text{Can-execute-plan}(b, \text{do}(b, \text{act}))$$

An agent can execute the plan of his performing an action if he can perform that action.

Axiom Execute2:

$$\text{True}(t, \text{Bel}(b, \text{True}(t, \text{occurs}(\text{pln})))) \Rightarrow \text{Can-execute-plan}(b, \text{pln})$$

An agent can execute a plan if he believes that it will occur.

Axiom Execute3:

$$\text{True}(t, \text{Can-execute-plan}(b, \text{pln1}))$$

$$\wedge \text{True}(t, \text{Bel}(b, \text{True}(t, \text{occurs}(\text{pln1}))) \wedge \text{length}(\text{pln1})=m$$

$$\Rightarrow \text{True}(t+m, \text{Can-execute-plan}(b, \text{pln2})))$$

$$\Rightarrow \text{True}(t, \text{Can-execute-plan}(b, \text{sequence}(\text{pln1}, \text{pln2})))$$

An agent can execute a sequence of two plans if he can execute the first and believes that as a result of executing the first, he will be able to execute the second.

Axiom Bel-Perform1:

$$\text{True}(t, \text{Bel-what-is}(a, \text{comb}(\text{safe}, t)))$$

$$\Rightarrow \text{True}(t, \text{Bel-how-to-perform}(a, \text{dial}(\text{comb}(\text{safe}, t))))$$

An agent believes that he can dial a safe if he believes that the combination of the safe is a particular value.

Axiom Can-perform1:

$$\text{True}(t, \text{Bel-how-to-perform}(a, \text{dial}(\text{comb}(\text{safe}, t)))) \wedge \text{True}(t, \text{At}(a, \text{safe}))$$

$$\Rightarrow \text{True}(t, \text{Can-perform}(a, \text{dial}(\text{comb}(\text{safe}, t))))$$

An agent can perform the action of dialing a safe if he believes that he can perform the action and is near the safe.

Axiom Can-perform2:

$$\text{True}(t, \text{Bel-what-is}(a, p)) \wedge \text{True}(t, \text{Friendly}(a, b))$$

$$\Rightarrow \text{True}(t, \text{Can-perform}(a, \text{tell-what-is}(b, p)))$$

An agent can tell a friend what something is if he himself believes that that thing has a particular value.

Axiom Can-perform3:

$$\text{True}(t, \text{Can-perform}(a, \text{do}(a, \text{request}(b, \text{act}))))$$

Agents can always perform request acts.

Now, the persistence rules:

Persistence Axiom1:

$$\text{True}(t, \text{At}(a, \text{object})) \wedge \neg \exists \text{ True}(t, \text{Occurs}(\text{Do}(b, \text{Move}(\text{obj}, \text{loc}))))$$

$$\vee \text{ True}(t, \text{Occurs}(\text{Do}(b, \text{Move}(a, \text{loc}))))$$

$$\wedge \neg \exists \text{ loc } \text{ True}(t, \text{Occurs}(\text{Do}(a, \text{Move-self}(\text{loc}))))$$

$\Rightarrow \text{True}(t+1, \text{At}(a, \text{obj}))$ If a is at the same location as some object obj , he will still be at the same location as the object at the next instant of time, unless in the intervening time, someone has moved the object or a , or a has moved himself to some other location.

Persistence Axiom2:

$$\text{True}(t, \text{Friendly}(a, b)) \Rightarrow \text{True}(t+1, \text{Friendly}(a, b))$$

Friends stay friends forever. An obvious idealization, but adequate for our purposes.

Persistence Axiom3:

$$\text{True}(t, \text{Equal}(\text{comb}(\text{Sf}), v)) \wedge \neg \exists a \text{ True}(t, \text{Occurs}(\text{Do}(a, \text{change-comb}(\text{Sf}))))$$

$$\Rightarrow \text{True}(t+1, \text{Equal}(\text{comb}(\text{Sf}), v))$$

Safe combinations stay the same unless someone performs the action of changing the combination.

Persistence Axiom4:

$$\text{True}(t, \text{Bel}(a, p)) \wedge \neg \exists b \text{ True}(t, \text{Trusts}(a, b)) \wedge \text{True}(t, \text{Occurs}(\text{do}(b, \text{tell}(a, \neg p)))) \Rightarrow$$

$$\text{True}(t+1, \text{Bel}(a, p))$$

We also add the following principles of belief:

Positive Introspection:

$$\text{True}(t, \text{Bel}(a, p)) \Rightarrow \text{True}(t, \text{Bel}(a, \text{True}(t, \text{Bel}(a, p))))$$

If an agent believes something, he believes that he believes it.

Consequential Closure:

$$\text{True}(t, \text{Bel}(a, p)) \wedge \text{True}(t, \text{Bel}(a, p \Rightarrow q)) \Rightarrow \text{Bel}(a, q)$$

Agents believe the logical consequences of their beliefs.

We must add necessitation on *all* the axioms, and state that the hearer and listener of a communicative action always believe that it has taken place when it has taken place.