A First-Order Axiomatization of the Surprise Birthday Present Problem: Preliminary Report

Leora Morgenstern
IBM T.J. Watson Research Center
Hawthorne, NY 10532
leora@steam.stanford.edu

Long version of paper presented at Commonsense-05: March 28, 2005

Abstract

This paper presents a solution in a first-order monotonic logic to a simplified version of the Surprise Birthday Present Problem, a challenge problem for the formal commonsense reasoning community. The problem concerns two siblings who wish to surprise their sister with a present for her birthday: the aim is to construct a theory that will support the desired inferences, not allow undesired inferences, and be sufficiently elaboration tolerant to support reasoning about problem variations. The theory presented in this paper includes the development of a possible-worlds analysis of the concept of surprise, and an extension to previous work on multiple-agent planning to handle joint planning and actions. We show that this theory can solve the original SBP as well as many of its variants.

1 Introduction

1.1 Problem Statement

This paper presents an initial solution in a first-order monotonic logic to a simplified version of the Surprise Birthday Present Problem [5], one of a set of challenge problems for the formal commonsense reasoning community. The problem concerns two siblings who wish to surprise their sister with a present for her birthday. The aim is to construct a theory that will support the desired inferences, not allow undesired inferences, and be sufficiently elaboration tolerant (as in [18]) to support reasoning about problem variations.

The problem is reproduced below, slightly condensed and paraphrased for the sake of brevity:

Alice and Bob want to surprise their sister Carol with a joint present for her birthday, two weeks from now. They therefore go into a closed room to decide on the present and to plan how they will buy it.

The problem is to determine that their plan will work. Variants on the problem include predicting that the plan will not work if Carol is also in the room; if the door is open and Carol is in the next room; if one of them tells Carol; if they do not consult together; if they cannot agree on a present; or if they wait until after Carol's birthday; as well as to predict that the plan will still work if Alice and Bob discuss the plan during a walk outside, or pass a hidden message, and whether they go together to buy the present or go separately.

The solution must satisfy the following constraints: first, the theory should not support the inference that nothing happens except the events enumerated in the plan (and these events' consequences); second, that the theory should not support the inference that Carol knows nothing except for statements true in all possible worlds.

1.2 The Approach

The Surprise Birthday Present Problem (SBP) is one of a set of mid-sized challenge problems proposed for the formal commonsense reasoning community. [1] These problems are larger than toy commonsense problems (Yale Shooting Problem [11], Suitcase Problem [16], Missionaries and Cannibals [18]) that have received much attention from formal AI researchers, but much smaller than large-scale efforts to formalize substantial chunks of commonsense knowledge, such as the HPKB [22] project. In contrast to toy problems, which eviscerate most interesting details of commonsense reasoning, and large-scale efforts whose size necessitates a shallow approach to formalizing knowledge, the aim is to construct a relatively deep formalization of the mid-sized problem domain.

The aim of constructing these mid-sized formalizations is threefold, as discussed in [20]. First, the goal is to create reusable core, reusable theories, or partial theories, of commonsense reasoning, as in [13] [12]. For example, in this paper, we develop some core definitions of expectation and surprise. Second, extending existing work into the mid-sized axiomatization tests the limits of existing theories: one either discovers that an existing theory is too brittle to be expanded to the demands of the non-toy formalization, or one invents methods to extend the existing theory. For example, this paper explores how the planning theory of [8] could be extended to joint plans. Third, analyzing a mid-sized problem could result in discovering new general representational issues and problems, which often would not be discovered when examining toy problems. (Typically such problems, once discovered, can be recast into toy problems.)

Even mid-sized problems turn out to be quite complicated, and many simplifications are necessary for formalization. (See, e.g., the simplifications used in [20], [26] for the formalization of the Egg Cracking Problem.) The SBP, as Davis has pointed out, concerns a variety of domains, including time, space, physics, knowledge, perception, naive psychology, multiple agents, and planning. Focusing on all these problems in depth would necessitate a large-scale, rather than a mid-sized axiomatization. It is also may lie beyond the capabilities of AI practitioners today. Each of these domains presents substantial challenges for formalization; moreover, integrating formalizations of separate domains usually presents various difficulties [4].

Instead, we focus on just two issues: formalizing the concept of surprise, and formalizing some concepts relating to joint plans. In this paper, we present preliminary work toward that goal. We first characterize the concept of surprise: an agent is surprised by a fact being true or an event happening if he previously did not expect it, but has subsequently found out about it. (This English paraphrase is a simplification, as we discuss in Section 2, where we provide a more accurate definition.) We then investigate the circumstances in which agents can successfully execute a joint plan. The formalization is an extension of the theory of [8], in which plans consist of a single agent making a request to single or multiple agents, each acting alone. The extension to scenarios in which multiple agents jointly form and execute plans presents several technical and conceptual issues.

To show that Alice's and Bob's joint plan results in Carol being surprised, we must posit, first, that in the starting situation, Carol has no expectation of receiving a gift on her birthday. We then specify the joint plan to purchase the present and give it to Carol. We show that this plan will succeed in Carol receiving the present. We furthermore show that Carol does not find out about the present during the execution of the plan. This is sufficient to entail the conclusion that Carol is surprised when she receives her birthday present.

We do not focus on the issue of physical proximity as it relates to overhearing a discussion: we simply introduce the *within_earshot* fluent to mean that one agent is in earshot of another. We ignore the issue of location entirely, not specifying that in order to give someone something or purchase something, one must be at a particular location. We make this choice because adding axioms about location would at least double the length of the already complex plan, and it seemed more productive to use the space to discuss other issues. It

[1] This set of problems can be found at http://www-formal.stanford.edu/leora/commonsense.

is of course understood that in an extended version of this paper, we would include such fluents.

This means, of course, that there is a large chunk of the SBP that we are not analyzing, and some variants that we cannot handle. However, what remains is still a very complex problem, as we demonstrate.

1.3 Logical Preliminaries

We will be using a sorted logic. *A* ranges over agents; *S* ranges over situations; *T* ranges over calendar-clock-times (e.g., January 5, 2005, 21:41), *E* ranges over events, *P* ranges over plans, *Q* range over fluents, and *X* ranges over objects. Other sorts will be introduced as needed. Variables are uppercase; constants are lowercase. In all statements, variables are assumed to be universally quantified unless otherwise specified.

We will be using the situation-based temporal logic of [8]. Time branches forward, but not back. The forward-branching structure represents the potential choices that agents can make; different choices correspond to different paths through the structure. Situations are ordered by the $<$ relation. Associated with each situation is a calendar-clock-time, also ordered by the $<$ relation.

Finite intervals are specified by their starting and ending situations. The predicate *holds* relates fluents and situations: *holds(S, Q)* means that the fluent *Q* is true in the situation *S*. We extend the notation so that *holds* can be used over intervals: *holds([S1,S2],Q)* $\Leftrightarrow \forall_{S \in (S1,S2)}$ *holds[S,Q]*

Events occur over intervals. *occurs(S1,S2, e)* means that event *E* occurs over the interval *[S1,S2]*.

2 Formalizing the concept of surprise

We formalize the concept of surprise as an unexpected event or fact. That is, *A* is surprised by *Q* at situation *S* if prior to *S*, *A* did not expect that *Q* would hold.

To formalize this concept, we must deal with two issues: First, formalizing the notion of expectation, and second, extending previous work on the interaction between time and knowledge (as well as between time and other knowledge-like operators), in the sense to be made precise below. Below, we discuss three knowledge-like operators: *Know*, *Believe*, and *Believe-likely*. Corresponding to these operators are three operators describing prediction: *Know-future*, *Believe-future*, and *Expect*.

Why the need for three epistemic/doxastic operators? First, we note that using only the *Know* operator limits the kind of surprise that can be expressed. Consider that one may be surprised by *Q* because one had no expectation that *Q*; but one may also, and in a somewhat stronger sense, be surprised by because one had the expectation that in fact ¬*Q* would hold. (We may distinguish these types of surprise as, respectively, weak surprise and strong surprise.) We can use *Know* to express weak but not strong surprise. For it is not possible for *A* to *Know* that ¬*Q* will hold at some time *T*, but for *Q* then to hold at *T*: knowledge implies truth. Although weak surprise is a sufficient concept for many situations (such as the SBP), we prefer to develop a theory that is capable of the fairly natural extension to the concept of strong surprise.

Second, the articulation and analysis of three separate concepts may help clear up a somewhat implicit muddle in the distinction between the commonsense concepts of belief and knowledge. One generally distinguishes between knowledge and belief by positing that knowledge, but not belief, implies truth [2]. However, there are also accounts in which belief is assumed to have a lesser degree of doxastic commitment than knowledge. That is, one is less sure of what one believes than of what one knows. The belief operator has different properties under these two accounts. In the former, certain axioms hold, such as KB.3, the second principle

[2] In terms of characterization, one generally does not *define* knowledge as true belief; one may, for example, attempt to define knowledge as justified true belief, as discussed in [10].

of positive introspection for belief, of [3] (if an agent believes something, he believes that he knows it) that do not hold in the latter; on the other hand, the latter concept is more suitable for theories of belief revision.

Indeed, for complex theories of multi-agent planning, all three concepts are important. If one agent *A1* reasons about another agent *A2*'s plan, *A1* must be able to distinguish between what *A2* knows to be true and what *A2* believes to be true: if *A2*'s beliefs are mistaken, A1 knows that A2's plan may not succeed. Likewise, *A1* knows that if *A2* believes that something is likely, but not necessarily true, *A2* may make some contingency plan that would not arise if *A2* believed something with certainty.

We therefore introduce three accessibility relations *K*, *B*, and *L*, relating, respectively, knowledge-accessible worlds, belief-accessible worlds, and likely-belief-accessible worlds. Intuitively:
K(A, S1, S2) holds if from what *A* knows to be true, *S2* is indistinguishable from *S1*;
B(A, S1, S2) holds if from what *A* believes to be true, *S2* is indistinguishable from *S1*;
L(A, S1, S2) holds if from what *A* believes to be likely, *S2* is indistinguishable from *S1*.

Definition 1 We then have the expected definitions:
$holds(S1, Know(A,Q)) \Leftrightarrow \forall_{S2}\ K(A,S1,S2) \Rightarrow holds(S2, Q)$
$holds(S1, Believe(A,Q)) \Leftrightarrow \forall_{S2}\ B(A,S1,S2) \Rightarrow holds(S2, Q)$
$holds(S1, Believe\text{-}likely(A,Q)) \Leftrightarrow \forall_{S2}\ L(A,S1,S2) \Rightarrow holds(S2, Q)$

As can be seen above, the definitions and axioms that we will have for knowledge, belief, and believing likely are often very similar. In this paper, we will frequently group related definitions and axioms together, to save space.

We specify that the *K* relation is reflexive and transitive, and that the *B* and *L* relations are symmetric and transitive, yielding an S4 logic of knowledge and weak S5 logics for belief and likely belief. This gives the usual axioms on epistemic and doxastic operators, as in [9].

The major differences between S4 and weak S5 are: (1) veridicality holds in S4 but not in weak S5; (2) negative introspection holds in weak S5 but not in S4. Indeed, we wish veridicality to hold only for *Know* but not for *Believe* or *Believe-likely*, since it is not the case that whenever *A* believes or believes-likely *Q*, that *Q* holds. Moreover, while negative introspection holds for *Believe* and *Believe-likely*, it doesn't hold for *Know*. Consider the case where *A* believes *Q* but doesn't know *Q*, because *Q* is in fact false. *A* does not know that he doesn't know *Q*, indeed, he believes that he knows *Q*.

Consequential closure holds in both S4 and weak S5 (and indeed, in any standard possible-worlds account of modal operators). In particular, consequential closure hpolds for *Believes-likely*. This is contrast to probabilistic models of likely belief, in which consquential closure cannot in general hold. Assuming consequential closure for *Believe-likely* has the advantage of facilitating temporal reasoning, specifically reasoning about causal chains and reasoning about the frame problem.

We further place the following restriction on these relations:

Axiom 1 $\{S2\ |L(A,S1,S2)\} \subseteq \{S2\ |B(A,S1,S2)\} \subseteq \{S2\ |K(A,S1,S2)\}$

To see that $L \subseteq B$, note that the greater an agent's degree of doxastic commitment, the fewer propositions to which he commits; therefore, the greater the number of worlds that are accessible to him. To see that $B \subseteq K$, note similarly that the truth requirement for knowledge, as opposed to belief, means that an agent can believe more propositions than he knows; since he commits, belief-wise, to more propositions than he commits, knowledge-wise, the set of knowledge-accessible worlds is larger than the set of belief accessible-worlds.

The subset relations on the accessibility relations correspond to the following axioms relating knowledge, belief, and likely belief:

Axiom 2 $holds(S, Know(A,Q)) \Rightarrow holds(S, Believe(A,Q))$

Axiom 3 *holds(S, Believe(A,Q))* $\Rightarrow$*holds(S, Believe-likely(A,Q))*

For many purposes — and in particular, for formalizing the notion of surprise — it is necessary to reason about the future. An agent may know that the fluent *Q* will hold at some future time; or believe that *Q* will hold; or believe that it is likely that *Q* will hold. To formalize these concepts, we will need to reason about the ways in which knowledge (resp. belief or believing likely) and time interact. Traditionally, theories of knowledge and time have formalized this interaction using some sort of *Results* function which maps a situation and the action performed in that situation to the situation resulting from the performance of that action [19]. This approach is useful for reasoning when one knows all the actions that one will perform, or at least a partial characterization of such actions [24]. However, we want to express an agent's ability to reason about the future even when he has little or no knowledge about the actions that will be performed. For example, we want to say that an agent can predict that the president will give the State of the Union address in January. Therefore, we need to express an agent's ability to reason about the future when that future is expressed not in terms of actions being performed but in terms of the passage of time or specific calendar dates.

To do this, we assume that the calendar-clock-time structure runs through all possible worlds, and that all agents always know (believe, believe likely) the date and time. [3] That is, they know (believe, believe likely) the calendar-clock-time of the situation they are in.

Axiom 4 *K(A,S1,S2)*$\Rightarrow$*time(S1) = time(S2)*

Due to the subset restriction on *L* and *B*, this means that we have as well that *B(A,S1,S2)*$\Rightarrow$*time(S1) = time(S2)* and *L(A,S1,S2)*$\Rightarrow$*time(S1) = time(S2)*.

We can now formalize the concept of an agent predicting the future. We say that an agent *A* knows (resp. believes, believes-likely) that *Q* will be true at some future time *T* if, for any knowledge accessible situation *S2*, *Q* will always be true at some situation *S3* later than *S2*, as long as *S3*'s time stamp is *T*. It does not matter what actions happen between *S2* and *S3*. All that concerns *A* is the time stamp of *S3*.

Note, below, that *Know-future* corresponds to *Know* and *Believe-future* corresponds to *Believe*, but that *Expect*, rather than *Believe-likely-future* corresponds to *Believe-likely*: *Expect* seems the closest English word for this concept and less awkward than *Believe-likely-future*.

Definition 2 *holds(S1, Know-future(resp. Bel-future, Expect)(A,Q,T))* $\Leftrightarrow$
$\forall_{S2,S3}$ *K(A,S1,S2) (resp. B(A,S1,S2), L(A,S1,S2))* $\wedge$*S2* $<$ *S3* $\wedge$*time(S3) = T* $\Rightarrow$*holds(S3, Q)*

We can extend this notation so that the third argument can be a time interval, in the expected way:

Definition 3 *holds(S1, Know-future(resp. Bel-future, Expect)(A,Q,[T1,T2]))* $\Leftrightarrow$
$\forall_{S2,S3,S4}$ *K(A,S1,S2) (resp. B(A,S1,S2), L(A,S1,S2))* $\wedge$*S2* $<$ *S3* $\wedge$*S3* $<$ *S4* $\wedge$*time(S3) = T1* $\wedge$*time(S4) = T2*
$\Rightarrow$*holds([S3,S4], Q)*

We further extend the definition, overloading the Know-future/Bel-future/Expect operators so that we can talk about predictions and expectations of event occurrences:

Definition 4 *holds(S1, Know-future(resp. Bel-future, Expect)(A,E, T1))* $\Leftrightarrow$
$\forall_{S2,S3}$ *K(A,S1,S2) (resp. B(A,S1,S2), L(A,S1,S2))* $\wedge$*S2* $<$ *S3* $\wedge$*time(S3) = T1* $\Rightarrow\exists_{S4}$ *occurs(S3,S4,E)*

[3] This assumption is indeed a theorem of the time structure set up in [7]. Note, however, the reliance of the proof on the S5 structure of the knowledge-accessibility relation, which is not present here.

We now consider the concept of surprise. We define *A* being surprised at *S1* by a fact *Q* being true at *S2* or an event *E* occurring starting at *S2*. First, consider a strawman version. It might seem reasonable to say that *A* is surprised if previous to *S2* he did not expect *Q* or *E* at *S2*. However, we wish to accommodate scenarios in which an agent expects *Q* or *E*, but then for some reason (such as obtaining information), changes his mind and no longer expects *Q* or *E*. Should it then happen that *Q* is true at *S2* or *E* occurs at *S2*, *A* would in fact be surprised. Therefore, we say that *A* is surprised if the following conditions hold:

• *S1* does not precede *S2*. • Any situation *S3* prior to *S2* in which *A* does not expect *Q* or *E* is followed by a later situation *S4*, still prior to *S2*, in which *A* does expect *Q* or *E*.

• In *S1*, *A* knows that *Q* has held or *E* has occurred starting at *S2*. • *S1* is the first situation for which this is true.

Note that the correct formalization of surprise entails the strawman version.

Since we overload the definition of surprise for both facts and events, two definitions follow.

Definition 5 *holds(S1, Surprise(A,Q, S2))* $\Leftrightarrow$

$$\begin{array}{l}
S1 \geq S2 \land \\
holds(S2,Q) \land \\
\forall_{S3<S2}\ holds(Expect(A,Q,time(S2)) \Rightarrow \\
\quad \exists_{S4}\ (S3 < S4 < S2 \land \neg\ holds(S4, Expect(A,Q,time(S2)))) \land \\
\forall_{S5} K(A,S1,S5) \Rightarrow \exists_{S6}\ S6 \leq S5 \land time(S6) = time(S2) \land holds(S6,Q) \land \\
\neg\exists_{S7}\ (S7 < S1 \land \forall_{S5}\ K(A,S7,S8) \Rightarrow \exists_{S9}\ S9 \leq S7 \land time(S9) = time(S2) \land holds(S9,Q))
\end{array}$$

By convention, we will say that *A* is surprised by an event *E* at the *beginning* of *E*'s occurrence.

Definition 6 *holds(S1, Surprise(A,E, S2))* $\Leftrightarrow$

$$\begin{array}{l}
S1 \geq S2 \land \\
\exists_{S2*}\ occurs(S2,S2*,E)) \land \\
\forall_{S3<S2}\ holds(Expect(A,E,time(S2)) \Rightarrow \\
\quad \exists_{S4}\ (S3 < S4 < S2 \land \neg\ holds(S4, Expect(A,E,time(S2)))) \land \\
\forall_{S5}\ K(A,S1,S5) \Rightarrow \exists_{S6,S6*}\ S6 \leq S5 \land time(S6) = time(S2) \land occurs(S6,S6*,E) \land \\
\neg\exists_{S7}\ (S7 < S1 \land \forall_{S5}\ K(A,S7,S8) \Rightarrow \exists_{S9,S9*}\ S9 \leq S7 \land time(S9) = time(S2) \land occurs(S9,S9*,E))
\end{array}$$

These definitions characterize the concept of weak surprise, as discussed above. To account for strong surprise, we must explicitly mention *A*'s expectation that $\neg$*Q* hold at *T*. We give the definition for fluents; the definition for events is analogous to weak surprise.

Definition 7 *holds(S1, Strong-surprise(A,Q, S2))* $\Leftrightarrow$

$$\begin{array}{l}
S1 \geq S2 \land \\
holds(S2,Q) \land \\
\forall_{S3<S2}\ \neg\ holds(Expect(A,\neg\ Q,time(S2)) \Rightarrow \\
\quad \exists_{S4}\ (S3 < S4 < S2 \land \neg\ holds(S4, Expect(A,\neg\ Q,time(S2)))) \land \\
\forall_{S5}\ K(A,S1,S5) \Rightarrow \exists_{S6}\ S6 \leq S5 \land time(S6) = time(S2) \land holds(S6,Q) \land \\
\neg\exists_{S7}\ (S7 < S1 \land \forall_{S5}\ K(A,S7,S8) \Rightarrow \exists_{S9}\ S9 \leq S7 \land time(S9) = time(S2) \land holds(S9,Q))
\end{array}$$

3 Joint plans

Central to the SBP is a complex notion of planning. Alice and Bob make a plan to buy Carol a gift and to give it to her, and subsequently execute that plan. The plan involves a variety of actions, performed by both

Alice and Bob; these actions must be coordinated properly. Moreover, in order for Alice and Bob to reason that their plan will succeed, they must know that they can and will both faithfully follow the agreed-upon plan. Reasoning about the success of the plan involves being able to reason about agents' constructing a joint plan, delegating and requesting, agreeing to requests, committing to plans, and reserving time to work on the plans to which they have committed.

We take as the basis of our work the theory of multi-agent planning developed in [8] and extend it to joint plans. That theory supports showing that certain multi-agent plans will succeed: in particular, plans in which one agent *requests* another agent, or requests a group of agents, by issuing a *broadcast request* to perform some plan. The theory has the following features: It is egalitarian in the sense that an agent cannot simply order other agents to drop their activities and immediately do what he asks. On the other hand, it is cooperative: every agent *reserves* blocks of time for every other agent and will work on a requesting agent's plan during a reserved time block if it does not interfere with another agent's plan. A fairly restrictive protocol specifies exactly when an agent *A* may *abandon* a requesting agent *A1*'s plan *P1* — specifically, when *A* has no way of continuing *P1* or when he is also committed to *A2*'s plan *P2*, and *P2* specifically forbids *A* from doing an action of *P1*. *A2* can specifically forbid *A* from doing an action if *A2 governs* that action. This ensures that *A* will not remain permanently committed to a plan that he cannot execute and that he will not do actions that interfere with other agents' plans.

A plan is specified in terms of two predicates, *succeed(P1, S1)* and *next_step(E, P1, S1, S2)*. *succeed(Pl, S1, S2)* is true if plan *P1*, started in situation *S1*, ends successfully in *S2*. *next_step(E, P1, S1, S2* is true if in *S2* action *E* is a possible next step of an instance of plan *P1* begun in *S1*. *next-step* is, essentially, the set of instructions for an agent to carry out a plan, specifying both the actions he needs to accomplish *P1* and the set of actions that he is permitted to do when, during the execution of *P1*, he momentarily turns his attention to work on another plan.

A proof in this theory of plan executability generally proceeds as follows: One shows that a plan *P* is executable by showing that in every unbounded-from-above *socially-possible* interval in which an agent *commits* to a plan, he *completes* that plan. Socially-possible intervals are those intervals in which all agents do what is requested of them to the extent possible.

An agent *completes* a plan over some interval if he *begins* the plan and *knows that the plan succeeds* over that interval. He *begins the plan* over some interval if he has begun it during that interval, and is still in the process of carrying out: that is, as long as the plan has not *terminated*, whenever he is at a *choice point* of deciding which action to perform, he knows of some action that is a *next-step* of the plan.

A plan is only *terminated* if it *succeeds* or if the *abandonment conditions* discussed above are satisfied.

The predicates corresponding to the italicized words above are discussed in detail in [8], where the complete set of axioms is given. The paper and a sample proof can be found, respectively, at www.cs.nyu.edu/cs/faculty/davise/elevator/axioms.ps and www.cs.nyu.edu/cs/faculty/davise/commplan-appb.pdf.

3.1 Extending the planning theory to agents acting together

In the theory of [8], agents, even in multiple-agent plans, always act alone. That is, a requesting agent may request a number of agents to do some actions, but agents never collaborate. For the SBP, however, we need to reason about joint plans in which agents collaborate and act together. There are several ways in which we must extend this theory to handle such plans:

1. Plan formation. In the original theory, a single requesting agent makes a request of one or more agents. For joint plans, there is no single agent who makes a request; rather, a group of agents jointly decide on a particular plan.

2. Reserving time blocks (related to the above point). In the original theory, all agents reserve time blocks for all other agents. That is how the requesting agent knows an agent will eventually have time to attend to his requests. Since for joint plans, there is no single requesting agent, it is unclear how time blocks will be reserved.

3. Joint actions. The original theory forces asynchronous action: only one agent may act at any particular time. Concurrency is possible in the sense that when *A1* is in the middle of performing some action, *A2* may start some other action. However, they cannot both start actions at the same time, and in particular, cannot both perform a single action at the same time. The most natural understanding of the SBP, however, is that Alice and Bob together give Carol her birthday present. That is, joint actions are necessary.

We discuss our approaches to these problems below:

3.1.1 Plan formation

There are several possible approaches. First, we could arbitrarily choose one of the agents in the plan as the requesting agent. This agent could formulate a plan in which he does his actions as they come up in the plan, and in which he requests the other agents to do their parts. This approach is problematic first because choosing which agent is the requesting agent is arbitrary; second, because it gives one agent considerable power over the others, for no good reason. A second possible approach would have each of the n agents in a plan request the other n-1 agents to do certain actions. The main drawback with this approach is the difficulty of coordinating all agents' actions among the n plans.

The third approach, which we adopt, is to posit a new entity, called a *joint plan entity* (JPE), that represents all the agents in the plan. A JPE is considered an agent; it is best thought of as similar to a corporate entity. The sort *J* ranges over joint plan entities. *members*(J) denotes the agents involved in the joint plan *J*. A particular joint plan associated with plan *Pi* is denoted J_{Pi}. $J \subset A$; in particular, this means that all axioms on agents apply to JPEs. We identify certain actors — those that are not joint plan entities — as individuals, denoted by the predicate *Individual(A)*. Actors that are joint plan entities are denoted by the predicate *JPE(A)*.

There are certain things that a JPE cannot do, such as accept plans from any agent including himself. Indeed, no agent is allowed to issue a request to a JPE.

Axiom 5 $\neg\exists_{S1,S2,A,J,P}$ *occurs(S1,S2,request(A,J,P))* $\vee$*accepts_request(P,A,J,S1)*

All joint plans have a similar structure. The joint plan entity starts the plan—and becomes active—with a broadcast request to all agents associated with the JPE, specifying the plan that the agents are to carry out; then the JPE waits. (Corporations issue orders, but do nothing else.) When the JPE's plan succeeds or is abandoned, the JPE ceases to be active. We introduce the predicate *active(J, S)*, indicating that JPE *J* is active in situation *S*.

Axiom 6 *holds(S, active(J))* $\Leftrightarrow$
occurs(S1, S2, broadcast_req(J,members(J),R)) $\wedge$
[S $\in$ *[S1,S2]* $\vee$
$[\exists_A$ *A* $\in$ *members(J)* $\wedge$*assignment(R,A) = P* $\wedge$*working_on(P,A,J,S2,S)]]*

It will be necessary to add the following axiom to the theory, stating that a JPE knows something if all agents in the entity know it:

Axiom 7 *holds(S, Know(J,Q))* $\Leftrightarrow$*[* $\forall_A$ *A* $\in$ *members(J)* $\Rightarrow$*holds(S, Know(A,Q))]*

It will also be necessary to modify the predicate *governs* which in the original theory ranges over an agent and an action. There will be times when joint plan entities govern many actions; however, we do not wish this governance to continue beyond the time that the joint plan is active. To express this, we need to add an extra situational argument to *governs*, and then specify that the joint plan governs actions only when it is active. (See Premise 8 for an example).

3.1.2 Reserving time blocks

The original theory posited that all agents reserve blocks of time for all other agents. (Each agent also reserves time for himself.) However, it is unrealistic to assume that all agents reserve blocks of time for all possible joint plan entities or even for all possible combinations of agents. Instead, we allow joint plan entities to cannibalize the reserved blocks of the plan members. That is, if *A1* and *A2* are members of some JPE *J*, some, but not all, reserved blocks of time that *A1* has reserved for *A2* will become reserved for J.

We alter the original theory as follows: The original theory has predicates *reserved(T, A1, A2)* meaning that time *T* is reserved by *A1* to work on a plan of *A2* and *reserved_block(A1,A2,T,D)* which is true iff all times between *T* and *T + D* are reserved by *A1* for *A2*.

We now call these predicates *init_reserved* and *init_reserved_block*, respectively. These apply only to individual agents; JPEs, which are created on the fly, do not initially reserve time for anyone else; nor does anyone assign time for them. We posit a function *allotment(S, A1, A2,* $\{J \mid A1, A2 \in members(J) \wedge active(J)\}$, *allotment-history(A1,A2,S))*. It takes as arguments a situation, 2 agents, all joint plans that are active in that situation and have those agents as members, and the *allotment history*. *allotment-history(A1,A2,S)* gives the sequence of blocks, starting at s0 (the starting situation of the world), and up to *S*, initially reserved by individual agent *A1* for individual agent *A2*, along with a record of who actually received the blocks: *A2* or some JPE with members *A1* and *A2*. The function *allotment* looks at the allotment history with respect to *A1* and *A2* as well as the set of currently active JPEs and determines to whom the block reserved by *A1* for *A2* should go.

We place some restrictions on this function. First, it is only defined if the agent of the second argument has originally (initially) reserved a block for the agent of the third argument. Second, when defined, the value must be either the agent of the third argument or a joint plan containing both the agents of the second and third arguments. It is assumed that the allotment function uses some sort of protocol, not specified here, to determine who next gets a block. We then define *reserved_block* using this allotment function; a block of time is reserved for whomever the allotment function decrees. Thus we have:

Axiom 8 *individual(A1)* $\wedge$*individual(A2)* $\Rightarrow$
(reserved_block(A1,A3,T,D) $\Leftrightarrow$*time(S) = T* $\wedge$
allotment(S,A1,A2,$\{$*J* $\mid$ *A1, A2* $\in$ *members(J)*$\wedge$*active(J)* $\}$*, allotment-history(A1,A2,S)) = A3)* $\wedge$
A3 = A2 $\vee$*[A1,A2* $\in$ *members(J)* $\wedge$*holds(S, active(J))* $\wedge$*A3 = J]*

JPEs reserve time only for themselves, and only when they are active.

Axiom 9 *reserved_block(J,A,T,D)* $\Rightarrow$*J=A*

Axiom 10 *holds(S, active(J))* $\Leftrightarrow$*[T = time(S)* $\Rightarrow$*reserved(T,J,J)]*

This scheme disturbs a concept in the original theory, that of *max_delay*, the maximum amount of time that can pass between successive blocks reserved for the same agent. One might think it possible to posit a constant *max_init_delay* and then determine, given the allotment function, what *max_delay* is. (*max_delay* is then not a constant, but a function, taking the same arguments as the allotment function.) However, if one allows joint plan entities to be created at will, and, in fact, allows for multiple joint plan entities to be created by the same groups of individual agents (the *committee curse*), one cannot necessarily put any upper bound on the value of *max_delay*. This corresponds to a truth in scheduling: if one keeps committing to new plans before one has finished existing plans, and keeps splitting one's time, there is no guarantee that one will accomplish *any* of the plans. One can mitigate this problem by placing some severe limitations on the activities of joint plan entities. First, we insist that there be no more than one active joint plan entity associated with each group of agents. (This corresponds, in a rough way, to the constraint in the original theory that a requesting agent cannot issue a request to another agent if that agent is already working on a request of his.)

Axiom 11 *members(J1) = members(J2)* $\Rightarrow\neg$*(holds(S, active(J1))* $\wedge$*holds(S, active(J2)))*

If there are more than a non-trivial number of agents in the world, there can still be very many joint plans active at any particular point: $2^n - (n+1)$ non-trivial joint plans, where n is the number of agents. [4] Completing plans is feasible, but will take much longer: this may be an important consideration if there are hard time constraints on the execution of the plan. (This is the case with the SBP.) That is, one can put an upper bound on the value returned by *max_delay*, but it may turn out too large to ensure successful results for planning with time constraints.

One could handle this problem by requiring that agents not accept any requests from joint plan entities if it knows that doing so would raise the maximum value of *max_delay* above some threshold. The extended theory would, in a sense, treat joint plan entities as second-class citizens: a cooperative agent would be required to accept plans from any individual agent, but not from any JPE. The trouble with this solution is that once an agent is not required to accept a plan, there is no guarantee at all of a plan succeeding.

We defer the general topic to future research. For this particular problem, we assume that there are only three agents: Alice, Bob, and Carol. Thus, there can be at most four non-trivial JPEs active at any time. We can assume some sort of straightforward protocol that cannibalizes time slices reserved for an agent in a round-robin fashion. This would mean that the *max_delay* is no more than five times the *max_init_delay*. One can make some simple assumptions on the times needed to peform actions and the maximum delay times that will ensure successful results, given the time available for the plan. (We make these assumptions in the premises.)

3.1.3 Joint actions

The original theory specified that agents act asynchronously, in order to avoid situations where we have to reason about the interaction of actions that are started at the same time. The most natural understanding of the SBP, however, is that Alice and Bob jointly give the present to Carol. We therefore extend the theory to handle joint actions while preserving as much of the spirit of the original theory as possible.

First, we still do not allow concurrency. *A1* and *A2* are still not allowed to perform different actions concurrently: we merely allow multiple agents to perform a single action.

Second, we set things up so that the disturbance in the asynchronous nature of the universe is minimized to the extent possible. Specifically, it is an axiom of the original theory that two agents never start an action or end an action at the same time. (The two go together, since agents are always active: as soon as an agent finishes one action, he starts another.) One exception is made for the start of time since all agents begin to be active then; this is handled by making that situation a specific exception to the axiom and positing that all agents start out by performing the action of waiting (varying amounts of time).

We employ a similar trick for joint actions. We introduce the notation *Do(*$\{$*A1,. . .,An*$\}$*, z)* to indicate agents

[4] There are 2^n possible joint plans, but the plan with no members and the *n* plans with only one member are of no interest.

A1 . . . *An* performing actional *z*. [5] The sort *G* ranges over groups of agents; $G \subset A$. Consider such a joint action. We enforce the following condition: when this joint action occurs over the interval *(S1, S2)*, all agents do not begin acting simultaneously. We do this by formalizing a joint action as occurring over an interval that has two segments: an initial segment where agents are waiting for the other agents to catch up, and a second segment where the performance of the joint action actually occurs. Following the performance of the joint action, all agents wait for varying lengths of time, as in the starting situation, to make sure that the performance of actions is once again asynchronous. It faciliates stating this if we assume that to perform an action that takes *max_action_time*, all agents must reserve a block of time equal in length to twice *max_action_time* $+n \cdot \epsilon$ and that the actual joint action occurs *max_action_time* into the interval.

Axiom 12 $\mathit{feasible}(do(\{A1 \ldots An\}, Z), S) \Leftrightarrow \exists_{S1,S2,\ldots,Sn} Sn > Sn-1 > \ldots S2 > S1 \wedge \mathit{time}(Sn) = \mathit{time}(S1) + \mathit{max\_action\_time} \wedge \forall_{i=1\ldots n}\ \mathit{reserved\_block}(\mathit{time}(Si),Ai,Aj,\ 2\cdot max\_action\_time\ + (i+1)\epsilon) \wedge \mathit{feasible}(do(Ai,Z), Sn)$

Axiom 13 $\mathit{occurs}(S1, S2, do(\{A1 \ldots An\}, Z)) \Rightarrow \forall_{i=1\ldots n}\ \mathit{occurs}(S2, S2{+}i \cdot \epsilon, do(Ai, wait))$

We also have to make the appropriate changes to Axiom A.4 in [8], to allow an exception for joint actions to the axiom of asynchrony:

Axiom A.4': $\mathit{choice}(A1,S1) \wedge \mathit{choice}(A2,S1) \Rightarrow A1 = A2 \vee \exists_{Z,S2<S1}\ \mathit{occurs}(S2,S1,do(\{A1,A2\},Z))$

A separate and different problem, which we do not solve in general here, is ensuring that multiple agents will simultaneously have some time block reserved for the JPE. For if they do not all have some identical block of time reserved for the JPE, they obviously cannot perform the joint action. There is no guarantee that this will in general happen.

For the SBP, however, we will need to deal with a joint action only once, when Alice and Bob jointly give Carol her birthday present on her birthday. We can handle the issue for this particular case by positing first, that Alice and Bob keep that day free; i.e., each reserves the entire day for herself/himself, and second, that in cases where agents reserve large blocks of time for themselves, the allotment function will be able to assign a block of time in which both of them can perform the joint action.

To accomplish this, we add an axiom stating that if there exists a JPE J_{Pi} with members $A1, \ldots, An$ and one of the actions in Pi is a joint action involving some subset of the members of J_{Pi}, and each of the agents in that subset has initially reserved an identical large block of time for himself, then the allotment function will assign an identical portion of this identical block of time to each of the agents in the subset to work on plan Pi. For our purposes, a reserved block of time is "large" if it is at least 24 hours, and the portion of the large block of time allotted to the agents in the joint plan is sufficient to perform an action.

Axiom 14 $(\mathit{members}(J_{Pi}) = AN \wedge AM \subseteq AN \wedge D \geq 24 \wedge A \in AM \Rightarrow [\ \mathit{init\_reserved\_block}(A,A,T,D)) \wedge \mathit{one\text{-}step}(E,Pi) \wedge E = do(AM,Z)$
$\Rightarrow \exists_{D1,D2}\ D1 < D2 < D3 \wedge D2\text{-}D1 \geq \mathit{max\_action\_time} \wedge \mathit{reserved\_block}(A,J_{Pi},T{+}D1,D2\text{-}D1)]$

This might disturb, though only slightly and temporarily, the round-robin allotment discussed previously. We could account for it by using a higher constant multiple of *max_init_delay* to get *max_delay*. We do this in the premises.

This ad hoc solution seems to hint at the larger solution. One can arrange for multiple agents to perform joint actions, when each of the agents has ceded large blocks of time in some way, whether each agent to himself or each agent to some corporate entity.

[5] An actional is an action unanchored by an agent; e.g., *give(A2, X)* is the actional of giving object *X* to agent *A2*. Agents perform actionals.

4 Proving that Alice and Bob's plan will work

In this section, we state Alice's and Bob's plan to give Carol a gift on her birthday, show that Alice and Bob will be able to execute the plan, and show that Carol will be surprised when she receives the gift.

We organize this section as follows: To avoid losing the reader as (s)he slogs through myriad axioms about ownership, purchase, the transfer of money, and all that agents know about these actions, we state the domain axioms at the end of this section. We first give the plan specification; then discuss the frame problem in the context of this theory, and then sketch the proof; this is followed by the statement of the problem premises and domain axioms. The reader is encouraged to flip to the axioms when reading through the plan specification and the proof sketch.

4.1 Plan Specification

As we have set up the problem, there are two plans: the JPE's plan to broadcast the request to Alice and Bob, and the joint plan that Alice and Bob carry out. (In a fuller treatment of this problem, there would be at least two more plans: one in which Alice and Bob decide to go into a closed room and one in which Alice and Bob come to a decision about which present to get and how to organize their plan to get it.)

A few remarks about these axioms. The predicate *first_opportunity(S2, AC, AR, S1, Q)* is true when *S2* is the first situation since *S1* when *AC* has reserved a block of time for *AR* and *Q* is true. This predicate is used when specifying plans: a plan specifies that an agent do some action at his first opportunity. The fluents that are used in statements of this sort are often quite complicated; therefore, they are usually abbreviated in the *next_step* specification and defined in subsequent axioms.

Specification of p1:

Plan p1 is specified as follows: At the first opportunity when Carol is not in earshot of Alice and Bob, the JPE broadcasts a request *r2* to Alice and Bob. At all other times, the JPE waits. (Recall that the JPE is an artificial entity created just for the formation of joint plans; its main function consists in broadcasting the request.)

Plan Spec Axiom 1 *next_step(E, p1, S1, S2)* $\Leftrightarrow$
action(E, J_{p1}*)* $\wedge$
first_opportunity(S2, J_{p1}*,* J_{p1}*, S1, p1_f)* $\Rightarrow$
instance(E, broadcast_req(J_{p1}*,* $\{$*alice,bob*$\}$*, r, S2)* $\wedge$
$\neg$*first_opportunity(S2,*J_{p1}*,*J_{p1}*), S1, p1_f)* $\Rightarrow$*action(E,* J_{p1}*) = wait*

p1_f is true when Carol is not in earshot of either Alice or Bob.

Plan Spec Axiom 2 *holds(S, p1_f)* $\Leftrightarrow\neg$*holds(S, in_earshot(carol, bob))* $\wedge\neg$*holds(S, in_earshot(carol, alice))*

The request that the JPE broadcasts to Alice and Bob is to perform the plan *p2*.

Plan Spec Axiom 3 *A = alice* $\vee$*A = bob* $\Rightarrow$*assignment(r,A) = p2*

p1 succeeds if Carol ultimately receives the gift on her birthday.

Plan Spec Axiom 4 *succeeds(p1,S1,SN)* $\Leftrightarrow$
$\exists_{SM,SN}$ *SM,SN* $\in$ *birthday(carol)* $\wedge$*occurs(SM, SN, do(carol, receive-gift))*

Specification of p2:

Plan *p2* is specified as follows: First Alice gives Bob \$10, earmarking it for the gift *xgift*. Then Bob gives himself \$10, earmarking it for *xgift*. (This step facilitates proving that this is indeed a joint gift: other formulations are possible but potentially more awkward. Also note that Alice needn't give Bob money before Bob earmarks his own money; such an order is not enforced by the plan.) Then Bob purchases *xgift*. Then Alice and Bob together give Carol the gift. The plan is formalized with the help of flags *p2_q1* ... *p2_q4* which trigger the events in the plan. These flags are specified in the premises below.

p2 must also specify the actions that are taken when Alice and Bob are not working for the JPE. This plan allows Alice and Bob to do almost any action, but places limitations on their abilities to spend money, give things, and talk. In particular, they cannot give money to anyone except for Bob unless they always have at least \$20 left or the money is going toward the purchase of the gift; they cannot give the gift to anyone but Carol, and not even to Carol until her birthday; and they are not allowed to tell anyone that there is a plan afoot which includes giving Carol the gift. The techniques used to represent informing an agent of relatively complex fluents is taken from [6].

Plan Spec Axiom 5 *next_step(E,p2,S1,S2)* $\Leftrightarrow$
action(E, alice) $\vee$*action(E, bob)* $\wedge$
p2_q1(S2,S1) $\Rightarrow$*E = do(alice, give-earmark-cash(bob, 10, xgift))* $\wedge$
p2_q2(S2,S1) $\Rightarrow$*E = do(bob, give-earmark-cash(bob, 10, xgift))* $\wedge$
p2_q3(S2,S1) $\Rightarrow$*E = do(bob, purchase(xgift))* $\wedge$
p2_q4(S2,S1) $\Rightarrow$*E = do(alice,bob, give(carol,xgift))* $\wedge$
(* Now the plan specifies the forbidden actions *)
(A1 = alice $\vee$*A1 = bob* $\vee$*A1 =* $\{$*alice,bob*$\}$*)* $\wedge$*(E = do(A1, give-cash(A2, N))* $\vee$*E = do(A1,purchase(X)))*
$\Rightarrow$*cash(A1, S2)* $\geq$ *N + 20* $\vee$*A = bob* $\vee$*X = xgift*
$\wedge$
$\neg$*time(S2)* $\in$ *birthday(Carol)* $\Rightarrow$*E* $\neq$ *do(A1, give(A3, xgift))*
$\wedge$
time(S2) $\in$ *birthday(Carol)* $\wedge$*E = do(A1, give(A3, xgift))* $\Rightarrow$*A3 = carol*
$\wedge$
$\neg\exists_{E1,P,A3,A4,X}$ *E = do(A1,Inform(A2,Q))* $\wedge$
[Holds(S,Q) $\Leftrightarrow\exists_{Si,Sj}$ *Si* $<$ *Sj* $\leq$ *S* $\wedge$*occurs(Si,Sj, request(A3,A4,P))* $\wedge$
one-step(E1,P) $\wedge$*E1 = do(A3,give(carol,X))*

Below is the specification for the plan flags for *p2*. *p2_q1* is set at the first opportunity that Alice has a reserved block of time for the JPE and also has at least \$10. (Plan *p2* above specifies that when that flag is set, Alice gives \$10 to Bob.) *p2_q2* is set at the first opportunity that Bob has a reserved block of time for the JPE and also has at least \$10. *p2_q3* is set at the first opportunity after both Alice and Bob have earmarked money for *xgift* that Bob has a reserved block of time and also has at least \$20. *p2_q4* is set at the first opportunity on Carol's birthday that Alice and Bob both have reserved blocks of time for the JPE and one of them has *xgift*.

Plan Spec Axiom 6 Fluents and flags:
first flag:
p2_q1(S,so) $\Leftrightarrow$*first_opportunity(S, alice,* J_{p1}*, ss, p2_q1_f)*
first flag fluent:
holds(S, p2_q1_f) $\Leftrightarrow$*cash(alice,S)* $\geq$ *\$10* $\wedge$*reserved_block(time(S),alice,* J_{p1}*, max_action_time)*
second flag:
p2_q2(S,so) $\Leftrightarrow$*first_opportunity(S, bob,* J_{p1}*, ss, p2_q2_f)*
second flag fluent:
holds(S, p2_q2_f) $\Leftrightarrow$*cash(bob,S)* $\geq$ *\$10* $\wedge$*reserved_block(time(S),bob,* J_{p1}*, max_action_time)*

third flag:
p2_q3(S,so) $\Leftrightarrow$first_opportunity(S, bob, J_{p1}, ss, p2_q3_f)
third flag fluent:
holds(S, p2_q3_f) $\Leftrightarrow\exists_{S1,S2,S3,S4}$ S1 $<$ S2 $<$ S $\wedge$S3 $<$ S4 $<$ S $\wedge$occurs(S1,S2, do(alice, give-earmark-cash(bob, 20, xgift))) $\wedge$
occurs(S3,S4, do(bob, give-earmark-cash(bob, 20, xgift))) $\wedge$cash(bob,S) $\geq$ 20 $\wedge$reserved_block(time(S),bob, J_{p1}, max_action_time)
fourth flag:
p2_q4(S,so) $\Leftrightarrow$first_opportunity(S, $\{$alice,bob$\}$, J_{p1}, ss, p2_q4_f)
fourth flag fluent:
holds(S, p2_q4_f) $\Leftrightarrow$
time(S) $\in$ birthday(carol) $\wedge$
holds(S, phys-possess(bob,xgift)) $\vee$holds(S, phys-possess(alice,xgift)) $\wedge$
reserved_block(time(S),$\{$alice,bob$\}$, J_{p1}, 2 $\cdot$ max_action_time + ϵ)

The success condition is simply that the steps in the plan have been completed in the appropriate order.

Plan Spec Axiom 7 *succeeds(p2,S1,SN) $\Leftrightarrow$*
$\exists_{S2,S3,S4,S5,S6,S7,S8,S9}$ S1 $<$ S2,S4,S6,S8 $\wedge$S2 $<$ S3 $\wedge$S4 $<$ S5 $\wedge$S3,S5 $<$ S6 $<$ S7 $<$ S8 $<$ S9 $\leq$ SN $\wedge$
occurs(S2,S3, do(alice, give-earmark-cash(bob, 10, xgift))) $\wedge$
occurs(S4,S5, do(bob, give-earmark-cash(bob, 10, xgift))) $\wedge$
occurs(S6,S7, do(bob, purchase(xgift))) $\wedge$
occurs(S8,S9, do($\{$alice,bob$\}$, give(carol, xgift)))

4.2 The Frame Problem in this Context

The frame problem [17] is the problem of determining which fluents stay the same in a changing world. When one specifies a theory, one generally specifies how actions change the world; for example, putting one block on top of another changes the location of the first block. However, in order to reason successfully, one must also reason about all the things that stay the same, such as the location of the second block.

Any solution to the frame problem works by making quite a lot of assumptions. One can make these assumptions within a monotonic logic or within a nonmonotonic logic. The primary advantage of a nonmonotonic logic is that one need not make these assumptions explicitly; the nonmonotonic reasoning mechanism does most of the work.

A common solution to the frame problem within a monotonic logic, popularized by Reiter [23], works by specifying *explanation closure* axioms: axioms that state the complete set of actions that can modify a fluent. A problem statement will typically have to state, explicitly or implicitly, that the actions that could modify a particular fluent do not in fact happen. (In the original situation calculus, the non-occurrence of actions not specified with the *Result* or *Do* function is implicit.)

Approaches to the problem in a nonmonotonic logic work similarly. Shanahan [25], for example, circumscribes the causal predicates of a theory (*Initiates*, *Terminates*, and *Releases*), essentially mirroring the effect of explanation closure, and circumscribes the occurrence predicate *Happens*, which entails, roughly, that as few actions as possible happen.

In this preliminary work, we proceed with a monotonic approach to the frame problem. There are three reasons for this choice. First, integrating theories is always time-consuming and often difficult: we believe that any attempt to integrate Shanahan's (or another) nonmonotonic solution with our theory of mult-agent and joint planning would overshadow any other aspect of the SBP. (In particular, the Event Calculus, on which

Shanhan's solution is based, provides an excellent representation for narratives, but is not so well suited for expressing plans.)

Second, the SPB problem description specifically states that a theory ought not entail that no actions happen other than the actions in the plan. But this is precisely what nonmonotonic solutions to the frame problem entail; that is how they work. Within a monotonic theory, one has a bit more latitude; it is easier to fine-tune things so that one does not wind up eliminating as many action occurrences. Thus, it is much easier to satisfy this constraint within a monotonic theory.

Third, the way the planning theory is set up, one anyway has to specify that certain actions are forbidden, namely, the actions that would interfere with the rest of the plan. These actions turn out to be remarkably similar to the sorts of actions one would have to explictly exclude from occurrence in a monotonic theory. For example, consider the actions that are prohibited to Alice and Bob in *p2*. They cannot spend down their money (before earmarking); they cannot give away the gift intended for Carol; they cannot tell anyone about their plan to give Carol a gift. These correspond to non-occurrence axioms stating that spending down money (below $20) never occurs; that Alice and Bob do not give away the gift; that Alice and Bob do not tell anyone about their plan. But they are not non-occurrence axioms: they are part of the plan specification.

This form of plan specification, therefore, has the potential to reduce the number of frame-problem-related assumptions one must make. One must still specify all explanation closure axioms; however, *if the only agents in the universe are the agents involved in the plan*, one may sometimes get away without extra non-occurrence axioms. One cannot get away without extra axioms if there are other agents in the universe, or if there are multiple plans, some of which don't specify non-occurrence. For example, for the SBP, one must specify that no occurrences of actions with potentially harmful effects happen during the time the JPE broadcasts the joint plan request to Alice and Bob.

The precise connection between plan specification and non-occurrence axioms as they relate to the frame problem is a subject for future research. In particular, we would be interested in determining whether non-monotonic techniques could aid in the formulation of the negative part of a plan specification, and in determining which conditions a requesting agent must govern.

4.3 Proof Sketch

We first show that plans *p1* and *p2* can be successfully executed, resulting in Carol receiving the gift, and then show that she will be surprised. In what follows below, we will frequently indicate, throughout the argument, whether a fact follows from the original theory (O), a lemma in the proof sketch (PS), or the extended theory (ET). [6]

The proof proceeds as follows: We begin by considering the second plan *p2*. Assume that between *ss* and *S1* J_{p1} issues a broadcast request to Alice and Bob to perform *p2*. Then, in any socially possible interval that includes *ss* and *S1*, both Alice and Bob accept the request to perform *p2*. Thus, both are committed to *p2* in *S1* (O).

Now consider the plan flags *p2_q1*, *p2_q2*, *p2_q3*, and *p2_q4*. We can show that Bob and Alice always know when these are true, and moreover, know when it is the first opportunity that a fluent holds (PS). For agents always know when they have reserved blocks of time (PS). Further, they know how much money they have, whether they own things, and know about the previous earmark-cash, purchase, and giving actions that they have performed (ET). We must show in addition that there will be such blocks of time available for Alice and Bob to perform their actions before Carol's birthday; and a block of time available on Carol's birthday for Alice and Bob to perform their joint action. This is a consequence of the problem premises specifying Alice

[6] The original theory and the proof sketch are available at www.cs.nyu.edu/faculty/davise/elevator/axioms.ps and www.cs.nyu.edu/faculty/davise/commplan-appb.pdf ; the extended theory refers to the development in this paper.

and Bob's free time, the maximum action time for doing actions, the maximum delay time during which agents can turn their attention to other plans, the length of time remaining until Carol's birthday, and (for the block of time available on Carol's birthday) the axiom (ET) on reserved blocks of time for joint plans.

We must show that
p2_qi(S2,S1) $\Rightarrow$*know_next_step(E, p2, alice,* J_{p1}*, S1)* $\Rightarrow$*E = do(alice, give-earmark-cash(bob, 10, xgift))*
(and similarly for the other plan steps).

For the first plan step, we must show that the action *E* is feasible in *S2* and that Alice knows that *E* is a next-step in the plan. We can show it is feasible in *S2* using the premises in the problem statement (i.e., Alice has $10), explanation closure axioms, the non-occurrence of events between *ss* and *S1*, and the conditions in the plan specification not allowing Alice to spend down below a certain amount of money.

We reason similarly to show that *p2_q2(S2,S1)* implies that Bob knows that the next step of *p2* is earmarking money for himself; feasibility is again shown using a combination of problem premises, non-occurrence of events, and explanation closure axioms. Similarly to show that *p2_q3(S2,S1)* implies that Bob knows that the next step of *p3* is purchasing the gift; and similarly to show that *p2_q4(S2,S1)* implies that both Alice and Bob know that the next step of *p2* is jointly giving the gift. For this last step, demonstrating feasibility appeals to requirements that the domain theory places upon joint giving: joint giving is possible only if all agents involved have earmarked money for the gift.

This will suffice to show that the predicate *begin_plan* (meaning, having begun and in the process of carrying out the plan, as long as the plan has not terminated) is true over any socially acceptable interval *[S1,Sz]*. Furthermore, we can show that the plan does not terminate before the final step of the plan has been performed. Termination can occur only if the plan succeeds or the plan has been abandoned; but neither of the abandonment conditions will be satisfied. For we have shown that it is always feasible for Alice and Bob to perform their steps in *p2*; and when, during *[S1,Sz]*, Alice and/or Bob are working on some other plan *p3* for some other agent, if they are requested to perform one of the forbidden actions, they will abandon *p3*, not *p2*, due to the fact that J_{p1} governs the forbidden actions. (O, ET)

We can also demonstrate certain properties of the situations in which the plan fluents first hold, using our premises on *max_action* and *max_delay*, and our axioms on allotment. In particular, we can show that the gift is purchased before Carol's birthday, and that there will be a first opportunity, on Carol's birthday, in which Alice and Bob both have allocated time for giving Carol her gift. (ET)

Finally, all agents know the actions that they have performed. Therefore, when the final step of the plan has been performed, Alice and Bob know it; therefore, they know the plan has succeeded. Thefore the plan completes, which means that the plan is executable. (O, ET)

Now let us turn our attention to *p1*. Since Alice and Bob are not in earshot of Carol in *ss*, they know that this is the case; therefore, J_{p1} knows it; therefore it knows that *p1_f* holds; further, it knows that *ss* is the first opportunity (since *ss*) when this is true. Furthermore it is always feasible to issue a broadcast request (O). Thus, in *ss*, J_{p1} knows the next step in plan *p1* and can perform it. Since it is always feasible to wait and no one governs the action of waiting (O), and this is known by all agents, we can show that once the request has been made, J_{p1} can continue to execute the plan *p1*.

In the proof sketch that *p2* was executable, we showed that Alice and Bob can reason that *p2* will successfully execute, and that Alice and Bob will jointly give *xgift* to Carol on her birthday. When this occurs, Alice and Bob will know that they have given the gift, and will therefore know that Carol has received the gift. Therefore, J_{p1} will know it. Thus the plan will complete and J_{p1} can successfully execute the plan.

Finally, we must show that Carol is surprised. Assume that *p1* executes over the interval *[ss,Sz]*. (Note that *p1* and *p2* complete at the same time.) Then there exists some situation *Sy* such that Alice and Bob give Carol the gift over *[Sy,Sz]*, where *[Sy,Sz]* is a subinterval of Carol's birthday.

Now we know from the problem premises that in *ss*, Carol does not expect to receive a gift on her birthday. We have as one of our explanation closure axioms that a person who does not expect *E* will come to expect that *E* will happen (prior to its occurrence) in one of only two ways: either by being informed that some plan that includes *E* is afoot, or by hearing a broadcast request to some agents of some plan that includes *E*. By hypothesis, Carol is not in earshot of Alice and Bob, and thus cannot hear the broadcast request. Moreover, no inform occurrences happen during the broadcast request. Furthermore, *p2*, which covers any time between the broadcast request and the giving of the gift, specifically forbids Alice and Bob telling anyone that anyone is working on a plan that includes giving Carol a gift on her birthday. Therefore, Carol will not be informed of the gift giving prior to her birthday.

By the definition of surprise, she will therefore be surprised when she receives her gift.

4.4 Domain Axioms

PREMISES: STARTING SITUATION

The only individual actors are Alice, Bob, and Carol:

Premise 1 $Individual(A) \Rightarrow A = alice \vee A = bob \vee A = carol$

In the starting situation ss Carol does not expect to receive a gift on her birthday.

Premise 2 $\neg[\ T \in birthday(carol) \Rightarrow holds(ss, Expect(carol, do(carol,receive\text{-}gift(carol)), T))]$

Alice and Bob each have at least $10.

Premise 3 $cash(alice,ss) \geq 10 \wedge cash(bob,ss) \geq 10$

The cost of xgift is $20.

Premise 4 $cost(gift1) = 20$

At the start, neither Alice, Bob, nor Carol owns the gift.

Premise 5 $\neg holds(ss, phys\text{-}possess(alice, xgift)) \wedge \neg holds(ss, phys\text{-}possess(bob, xgift)) \wedge \neg holds(ss, phys\text{-}possess(carol, xgift))$

At the start, Carol is not in earshot of Alice or Bob.

Premise 6 $\neg holds(ss, in\_earshot(carol, bob)) \wedge \neg holds(S, in\_earshot(carol, alice))$

We need some housekeeping axioms concerning reserved blocks of time, the length of time to do actions, the amount of time until Carol's birthday, etc. We assume units of one hour.

Premise 7 Housekeeping axioms:
$T = start(birthday(carol)) \Rightarrow init\_reserved\_block(T, alice, alice, 24)$
$T = start(birthday(carol)) \Rightarrow init\_reserved\_block(T, bob, bob, 24)$
$max\_action\_time = .5$
$max\_init\_delay = 2$
$max\_delay = 20$
$start(birthday(carol)) \geq time(ss) + 312$ (at least 13 full days till the birthday starts)

The joint plan entity J_{p1}, while it is active, governs the following actions of Alice and Bob: their spending down to below $20; their giving anyone *xgift*, and their telling anyone about a plan to give Carol a gift. Note that the governance axioms are very similar to the specification of the forbidden actions in *p2*.

Premise 8 What J_{p1} governs:
J_{p1} governs spending down below $20 unless purchasing the gift:
active(J_{p1}*,S)* $\Rightarrow$
  (A1 = alice $\vee$*A1 = bob* $\vee$*A1 =* $\{$*alice,bob*$\}$*)* $\wedge$
(E = do(A1, give-cash(A2, N)) $\wedge$*A2* $\neq$ *bob)* $\vee$*(E = do(A1, purchase(X))* $\wedge$*X* $\neq$ *xgift))* $\wedge$
cash(A1, S) $<$ *N + 20*
  $\Rightarrow$*governs(*J_{p1}*, E, S)*
J_{p1} governs giving the gift before Carol's birthday, and to anyone but Carol on her birthday
$\neg$*time(S)* $\in$ *birthday(Carol)* $\wedge$*E = do(A1, give(A3, xgift))* $\vee$
time(S) $\in$ *birthday(Carol)* $\wedge$*E = do(A1, give(A3, xgift))* $\wedge$*A3* $\neq$ *carol*
  $\Rightarrow$*governs(*J_{p1}*, E, S)*
J_{p1} governs telling anyone about working on a plan that involves giving Carol a present
$\exists_{E1,P,A3,A4,X}$ *(E1 = do(A3,give(carol,X))* $\wedge$*one-step(E1,P)* $\wedge$*E= do(A1,Inform(A2,Q))* $\wedge$
  [Holds(S1,Q) $\Leftrightarrow\exists_{Si,Sj}$ *Si* $<$ *Sj* $\leq$ *S1* $\wedge$*occurs(Si,Sj, request(A3,A4,P))]*
  $\Rightarrow$*governs(*J_{p1} *, E, S)*

PRECONDITIONS ON ACTIONS:

You can give cash to someone if you own at least that amount of cash:

Axiom 15 *feasible(do(A1,give-cash(A2,N)),S)* $\Leftrightarrow$*choice(A1,S)* $\wedge$*cash(A1,S)* $\geq$ *N*

That same precondition holds for giving cash to someone, earmarked for a particular purpose. A richer theory would require that this concept entail some notion of intention; this is deferred to future work.

Axiom 16 *feasible(do(A1,give-earmark-cash(A2,N,X)),S)* $\Leftrightarrow$*choice(A1,S)* $\wedge$*cash(A1,S)* $\geq$ *N*

You can give an object to someone if you physically possess that object:

Axiom 17 *feasible(do(A1,give(A2,X)),S)* $\Leftrightarrow$*choice(A1,S)* $\wedge$*holds(S, phys-possess(A1,X))*

You can buy something as long as you have sufficient cash:

Axiom 18 *feasible(do(A,purchase(X)),S)* $\Leftrightarrow$*choice(A,S)* $\wedge$*cost(X) = N* $\wedge$*cash(A,S)* $\geq$ *N*

Two agents can jointly give an object to a third agent as long as the following conditions hold:
— one of them physically possesses the object
— both of them have contributed money earmarked toward the object (before the giving of the object, not necessarily before the purchase of the object). The amount of money relative to the cost of the object isn't specified; any contribution is sufficient. However, the amount of each agent's contribution must be less than the cost of the object; otherwise, the others' earmarking doesn't count.
— both of them have reserved appropriate blocks of time as discussed above.

Axiom 19 *feasible(do(*$\{$*A1,A2*$\}$*, give(A3, X)), S)* $\Leftrightarrow$
(holds(S, phys-possess(A1,X)) $\vee$*holds(S,phys-possess(A2,X)))* $\wedge$
$\exists_{S2,S3,S4,S5,S6,S7,A4,N1,N2,D,Sj}$ *S2* $<$ *S3* $\wedge$*S4* $<$ *S5* $\wedge$*S6* $<$ *S7* $\wedge$*S3,S5,S7* $<$ *S* $\wedge$*(A4=A1* $\vee$*A=A2)* $\wedge$*N1* $<$ *cost(X)* $\wedge$*N2* $<$ *cost(X)* $\wedge$

$\mathit{occurs(S2,S3, do(A1, give\text{-}earmark\text{-}cash(A4,N1,X)))} \wedge \mathit{occurs(S4,S5, do(A2, give\text{-}earmark\text{-}cash(A4,N2,X)))}$
$\wedge$
$\mathit{occurs(S6,S7, do(A4,purchase(X)))} \wedge$
$\mathit{time(Sj) = time(S) + D} \wedge \mathit{choice(A1, S)} \wedge \mathit{occurs(S, Sj, do(A1, wait))} \wedge \mathit{choice(A2, Sj)}$

CAUSAL AXIOMS:

If one agent gives an object to a second, the first agent no longer has it, and the second does.

Axiom 20 $\mathit{occurs(S1,S2, do(A1, give(A2,X)))} \Rightarrow \mathit{holds(S2, Phys\text{-}possess(A2,X))} \wedge \neg \mathit{holds(S2, Phys\text{-}possess(A1,X))}$

The transfer of money works similarly:

Axiom 21 $\mathit{occurs(S1,S2, do(A1, give\text{-}cash(A2,N)))} \Rightarrow \mathit{(cash(A1,S2) = cash(A1,S1) - N)} \wedge \mathit{(cash(A2,S2) = cash(A2,S1) + N)}$

Purchasing an object results in an agent possessing the object but having less money.

Axiom 22 $\mathit{occurs(S1,S2, do(A, purchase(X)))} \Rightarrow \mathit{holds(S1, phys\text{-}possess(A,X))} \wedge \mathit{cash(A,S2) = cash(A,S1) - cost(X)}$

If someone tells you something, you will believe it.

Axiom 23 $\mathit{occurs(S1,S2,Inform(A1,A2,Q))} \Rightarrow \mathit{holds(S2,Believe(A2,Q))}$

If *A1* overhears *A3* requesting *A2* to do some plan, he will subsequently know that *A2* has accepted the request to perform that plan. (Note that *A1* can *know* that the plan will be accepted because agents always accept requests to do plans, as long as they do not have an outstanding request from the requesting agent. But the requesting agent is constrained from making such a request.)

Axiom 24 $\mathit{occurs(S1,S2, request(A3,A2,P))} \wedge \mathit{holds(S1,in\_earshot(A1,A2))}$
$\Rightarrow$
$\mathit{[\,[holds(S,Q)} \Leftrightarrow$
$\mathit{time(S1) = time(S2)} \Rightarrow \mathit{accepts\_req(P,A2,A3,S)]}$
$\Rightarrow \mathit{holds(S2, Knows(A1,Q))\,]}$

The axiom above will be used together with the following. If *A1* knows or even just believes that *A2* has accepted a request from *A3* to perform *P*, and one of the steps of *P* is some action *E*, then he will expect *E* to be performed at some time in the future. This is an expectation rather than knowledge of some future event, because *A1* does not necessarily know that all the circumstances that are crucial for the success of *P* to actually hold.

Axiom 25 $\mathit{holds(S2, Believe(A1,Q))} \wedge$
$\mathit{holds(S,Q)} \Leftrightarrow \mathit{(K(A1,S2,S)} \Rightarrow \mathit{accept\_req(P,A2,A3,S)}$
$\wedge \mathit{one\text{-}step(E,P)]} \Rightarrow$
$\exists_T\ T > \mathit{time(S2)} \wedge \mathit{holds(S2,Expect(A1,E,T))}$

RELATIONS BETWEEN ACTIONS:

If one has given cash to someone earmarked for some purpose, one has certainly given them cash.

Axiom 26 $\mathit{occurs(S1,S2, do(A1, give\text{-}earmark\text{-}cash(A2, N, X)))} \Rightarrow \mathit{occurs(S1,S2, do(A1, give\text{-}cash(A2,N)))}$

Giving entails receiving.

Axiom 27 *occurs(S1,S2, do(A1, give(A2,S)))* $\Rightarrow$*occurs(S1,S2, do(A2, receive(A1,X)))*

Definition of receiving some present from someone:

Definition 8 *occurs(S1,S2, do(A1, receive-gift))* $\Leftrightarrow\exists_{A2,X}$ *occurs(S1,S2, do(A1,receive(A2, X)))*

KNOWLEDGE AXIOMS:

Axiom 28 *In* ss*, Alice and Bob know all the premises. (Premises, not axioms; since axioms are true in all possible worlds, these are known in any theory based on a possible-worlds semantics for knowledge.)*

This means, for example, that Alice and Bob know in *ss* that Carol is not in earshot and that the gift costs $20.

Agents always know when it's someone's birthday:

Axiom 29 *time(S1)* $\in$ *birthday(A1)* $\wedge$*K(A2,S1,S2)* $\Rightarrow$*time(S2)* $\in$ *birthday(A1)*

Agents always know how much money they have:

Axiom 30 *Cash(A,S1) = N* $\Rightarrow$*(K(A,S1,S2)* $\Rightarrow$*Cash(A,S2) = N)*

Agents always know when they have been involved in a giving action, earmarking money action, or purchasing action:

Axiom 31 *Knowledge of actions:*
occurs(S1,S2, do(A, purchase(X))) $\wedge$*K(A, S2,S2A)* $\Rightarrow$
  $\exists_{S1A}$ *K(A, S1,S1A)* $\wedge$*occurs(S1A, S2A, do(A, purchase(X)))*
occurs(S1,S2, do(A1, give(A2, X))) $\Rightarrow$
  (K(A1, S2,S2A) $\Rightarrow\exists_{S1A}$ *K(A1, S1,S1A)* $\wedge$*occurs(S1A, S2A, do(A1, give(A2,X)))* $\wedge$
  (K(A2, S2,S2A) $\Rightarrow\exists_{S1A}$ *K(A2, S1,S1A)* $\wedge$*occurs(S1A, S2A, do(A1, give(A2,X)))*
occurs(S1,S2, do(A1, give-cash(A2, N))) $\Rightarrow$
  (K(A1, S2,S2A) $\Rightarrow\exists_{S1A}$ *K(A1, S1,S1A)* $\wedge$*occurs(S1A, S2A, do(A1, give(A2,N)))* $\wedge$
  (K(A2, S2,S2A) $\Rightarrow\exists_{S1A}$ *K(A2, S1,S1A)* $\wedge$*occurs(S1A, S2A, do(A1, give(A2,N)))*
occurs(S1,S2, do(A1, give-earmark-cash(A2, N,X))) $\Rightarrow$
  (K(A1, S2,S2A) $\Rightarrow\exists_{S1A}$ *K(A1, S1,S1A)* $\wedge$*occurs(S1A, S2A, do(A1, give-earmark-cash(A2,N,X)))* $\wedge$
  (K(A2, S2,S2A) $\Rightarrow\exists_{S1A}$ *K(A2, S1,S1A)* $\wedge$*occurs(S1A, S2A, do(A1, give-earmark-cash(A2,N,X)))*

EXPLANATION CLOSURE AXIOMS

The only way to have less money is to give it to someone or purchase an item:

Axiom 32 *S1* $<$ *S2* $\wedge$*cash(A1,S1) = N1 and cash(A1,S2) = N2* $\wedge$*N1* $<$ *N2*
$\Rightarrow$*(*$\exists_{X,S3,S4}$ *S1* $\leq$ *S3* $<$ *S4* $\leq$ *S2* $\wedge$*occurs(S3,S4, do(A1, purchase(X)))* $\wedge$*cost(X) = N1 -N2)* $\vee$
($\exists_{N3,A2,X,S3,S4}$ *S1* $\leq$ *S3* $<$ *S4* $\leq$ *S2* $\wedge$*(occurs(S3, S4, do(A1, give-cash(A2,N3)))* $\vee$*occurs(S3, S4, do(A1, give-earmark-cash(A2,N3,X))))*

The only way to lose possession of an item is to give it to someone. (For this preliminary version, we are not considering the pre-purchase owner of an item; this was done in order to minimize the total number of agents.)

Axiom 33 *S1* $<$ *S2* $\wedge$*holds(S1, phys-possess(A1,X))* $\wedge\neg$*holds(S2, phys-possess(A1,X))*
$\Rightarrow\exists_{A2,S3,S4}$ *S1* $\leq$ *S3* $<$ *S4* $\leq$ *S2* $\wedge$*occurs(S3,S4, do(A1, give(A2,X)))*

The only way to gain possession of an item is to get it from someone or to purchase it:

Axiom 34 *S1* $<$ *S2* $\wedge\neg$*holds(S1, phys-possess(A1,X))* $\wedge$*holds(S2, phys-possess(A1,X))*
$\Rightarrow$*[* $\exists_{A2,S3,S4}$ *S1* $\leq$ *S3* $<$ *S4* $\leq$ *S2* $\wedge$*occurs(S3,S4, do(A2, give(A1,X)))* $\vee$
$\exists_{X,S3,S4}$ *S1* $\leq$ *S3* $<$ *S4* $\leq$ *S2* $\wedge$*occurs(S3,S4, do(A1, purchase(X)))]*

If someone does not expect an action to happen, he will revise his expectations only if he find outs that there is a plan afoot that includes the action. In this formalization, we assume that there are only two ways for this to happen: one can overhear such a plan request being issued, or one can be told that such a plan request has been issued.

Axiom 35 $\neg$*holds(S1, Expect(A1,E,T))* $\wedge$*S2* $>$ *S1* $\wedge$*holds(S2,Expect(A1,E,T))*
$\Rightarrow$
$\exists_{Ax,Ay,Az,Sx,Sy,X}$ *S1* $\leq$ *Sx* $<$ *Sy* $\leq$ *S2* $\wedge$
one-step(E,P) $\wedge$*E=do(Ay, give(carol,X))* $\wedge$
[holds(Sx, in_earshot(A,Ay)) $\wedge$*[occurs(Sx,Sy,request(Ax,Ay,P))]* $\vee$
occurs(Sx,Sy,do(Az,inform(A,Q))) $\wedge$
[holds(Q,S) $\Leftrightarrow$
$\exists_{Si,Sj}$ *Si* $<$ *Sj* $\leq$ *S* $\wedge$*occurs(Si,Sj,request(Ax,Ay,P))]*

NON-OCCURRENCE AXIOM

Alice and Bob just wait while J_{p1} is broadcasting the request to do *p2*:

Axiom 36 *occurs(S1,S2, broadcast*$_req(J_{p1}, \{$*alice,bob*$\}$*, R)))* $\wedge$
(A1= alice $\vee$*A1 = bob)* $\Rightarrow$*assignment(R,A1) = p2*
$\Rightarrow$*occurs(S1,S2, do(alice, wait))* $\wedge$*occurs(S1,S2, do(bob, wait)))*

5 Problem Variants

Any formalization of a commonsense reasoning problem must be judged, at least in part, on how easily that theory can be extended to handle problem variants, that is, by how elaboration tolerant [18] the theory is. Below, we discuss the variants that the theory presented in this paper can currently handle, and how we might extend the theory to handle other variants.

Since we make no reference to location and have no theory of spaces or rooms, we clearly cannot handle certain variants: those where Carol is in the room where Alice and Bob are doing the planning, or where Carol is in the next room and the door is open. We likewise cannot handle the cases where Alice and Bob formulate their plan during a walk outside or pass a hidden message.

We can, however, handle an important subset of the variants. First, we can handle the variant when Carol is in earshot of Alice and Bob. We have an axiom stating that if an agent overhears someone requesting a plan, he knows that it will be accepted. Moreover, this agent will expect that any event that is a step of the plan will occur. Thus, if Carol hears the JPE broadcasting its request to Alice and Bob, she will expect to get a present. Similarly, the theory can handle the variant in which someone tells her that some agents are working on a plan that includes giving her a gift.

We can handle, in part, the variant in which Alice and Bob cannot agree on a present. In such a case, presumably there will be no joint plan entity, which would mean there is no earmarking of cash, no purchase, no joint gift. As yet, we have not sufficiently formalized the concept of joint plan entity to express or entail what it means when Alice and Bob cannot agree on a joint plan. We do not formalize the stage where the plan is discussed; we only formalize the stage where the plan is laid out in detail. For similar reasons, we cannot entirely handle the variants where Alice and Bob do not consult together. We can, however, certainly show that if Alice or Bob purchases the gift alone, without the other having earmarked money toward that purpose, that it does not count as a joint gift.

It is also possible, using this theory, to reason about a variant in which Alice and Bob wait to give Carol her gift until after Carol's birthday. One cannot reason about this particular circumstance if, in the starting situation, the joint plan entity requests Alice and Bob to perform *p2*, since the *next_step* specification of *p2* requires that the gift be given on Carol's birthday. However, one could formulate a similar plan *p3* in which Alice and Bob give the gift at the first possible opportunity after 12:01 AM on Carol's birthday, and one could alter the axioms on allotment and reserved blocks so that it is not necessarily the case that Alice and Bob are able to give the gift on Carol's birthday. Then although one could show that once Carol gets the gift, she is surprised, one could also demonstrate that it is possible that it is not on her birthday that Carol is surprised.

The theory can easily handle situations in which Alice and Bob jointly buy the gift. One can either specify the plan to include a joint purchasing action, or specify that the purchasing action may be done either jointly or singly, by either Alice or Bob.

6 Conclusion

This paper presents the results of the first phase of our work in constructing a first-order axiomatization for a simplified version of the Surprise Birthday Present Problem, one of a set of challenge problems for the commonsense reasoning community. Thus far, our results include the development of a possible-worlds semantics for the concepts of surprise, and the extension of a first-order theory for communication and planning to handle joint planning and action. We defer to future work proving that this extension shares the properties (such as consistency) of the original theory presented in [8].

We have demonstrated that this theory, together with some rudimentary axioms on giving, transferring money, and purchasing, suffices to demonstrate the goal of the SBP — showing that Carol is surprised when she receives her gift — and that we can handle many of the listed variants. In addition, the axiomatization satisfies the constraints set forth in the problem: the theory does not entail that Carol knows nothing of consequence, and does not entail that nothing happens except for the actions in the plan.

The second phase of this project will include developing the axiomatization for a less restricted version of the SBP. There are two major gaps in the axiomatization as it now stands: the lack of a theory of perception, and in particular, an integrated theory of perception and knowledge, and the lack of an account of how agents come to decide on a collaborative plan. The work of [2] on inferring ignorance from what is not perceived is, not surprisingly, particularly relevant to the first issue. Existing work on multi-agent negotiation [15] and on intentionality (joint or otherwise) [1] [14] may be relevant to the second.

Mid-sized axiomatizations of this sort are not common in the AI logicist community. This has hampered the development of a set of criteria for evaluation: As Nagel ([21] has argued, it is difficult to establish evaluative criteria before there is a critical mass of work in a particular area. Nevertheless, we can tentatively suggest some criteria, as in [20]. One can evaluate how well an axiomatization solves a challenge problem by how well it handles the problem itself and its variants. On this scale, this preliminary axiomatization seems solid: it can handle all variants within the intended scope of the axiomatization. One can evaluate how useful an axiomatization is by the generality and reusability of the core theories that it embodies. In this case, the theory of expectation and surprise is entirely general, and ought to be easily reusable. The theory of joint plans extends an existing theory of communication and multi-agent planning. The existing theory itself is much broader than most theories of multi-agent planning, and the extensions developed in this paper make it still more general.

Beside these criteria, there are some intangible benefits of doing this sort of mid-sized axiomatization that can transcend the criteria discussed above. Deep, narrow research into toy-sized problems rarely leads one to consider the many aspects of reasoning that simultaneously permeate even everyday commonsense reasoning problems: as we have seen, even a simplified version of the SBP involves the need to coordinate joint actions, to make sure that one will have time available, to earmark money to ensure that one has really contributed to a joint gift, and to reason about how all of these interact. There is ample opportunity, when working on large-scale formalizations, for such considerations to enter into one's consciousness, but virtually no time to spend thinking about any of the subtleties; there is too much to be done and never enough time to do it.

This is the level of formalization that presents the opportunity to reason about the multitude of ways in which various pieces of commonsense knowledge interact, and permits the time to develop one's theories as fully and as deeply as one can or would wish. The ultimate goal of such exercises may be the development of a sizable body of commonsense reasoning that can be used to solve larger, more serious problems, but even before that goal is met, the process itself captures some of the spirit of the original AI logicist enterprise.

References

[1] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2–3):263–309, 1990.

[2] Ernest Davis. Inferring ignorance from the locality of visual perception. In *Proceedings of the 7th National Conference on Artificial Intelligence(AAAI-1988)*, pages 786–790. AAAI Press/MIT Press, 1988.

[3] Ernest Davis. *Representations of Commonsense Knowledge*. Morgan Kaufmann, San Francisco, 1990.

[4] Ernest Davis. The naive physics perplex. *AI Magazine*, 19(3):51–79, 1998.

[5] Ernest Davis. The surprise birthday present problem, 2001. http://www-formal.stanford.edu/leora/commonsense/birthday.

[6] Ernest Davis. A first-order theory of communicating first-order formulas. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference (KR2004)*, pages 235–245, 2004.

[7] Ernest Davis. Knowledge and communication: A first-order theory. *submitted to Artificial Intelligence*, 2004.

[8] Ernest Davis and Leora Morgenstern. A first-order theory of communication and multi-agent plans. *Journal of Logic and Computation, to appear*, 2005.

[9] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge*. The MIT Press, Cambridge, Massachusetts, 1995.

[10] Edmund L. Gettier. Is justified true belief knowledge? *Analysis*, 23:121–123, 1963.

[11] Steve Hanks and Drew V. McDermott. Nonmonotonic logic and temporal projection. *Artificial Intelligence*, 33(3):379–412, 1987.

[12] Patrick Hayes. Naive physics I: Ontology for liquids. In Jerry Hobbs and Robert Moore, editors, *Formal Theories of the Commonsense World*, pages 71–107. Ablex, Norwood, New Jersey, 1975.

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

[14] Jerry R. Hobbs. Artificial intelligence and collective intentionality (a reply to john searle). In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication*, pages 445–460. MIT Press, Cambridge, Massachusetts, 1990.

[15] Sarit Kraus. Negotiation and cooperation in multi-agent environments. *Artificial Intelligence*, 94(1-2):79–97, 1997.

[16] Fangzhen Lin. Embracing causality in specifying the indirect effects of action. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI-95*, pages 1985–1993, 1995.

[17] John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, Edinburgh, 1969.

[18] John L. McCarthy. Elaboration tolerance. In *Working Papers of the Fourth International Symposium on Logical Formalizations of Commonsense Reasoning, Common Sense 98*, 1998.

[19] Robert C. Moore. Reasoning about knowledge and action, 1980.

[20] Leora Morgenstern. Mid-sized axiomatizations of commonsense problems: A case study in egg-cracking. *Studia Logica*, 67(3):353–384, 2001.

[21] Ernest Nagel. *The Structure of Science*. Harcourt, Brace, and Co., New York, 1961.

[22] Adam Pease, Vinay K. Chaudhri, Fritz Lehmann, and Adam Farquhar. Practical knowledge representation and the DARPA high performance knowledge bases project. In Anthony G. Cohn, Fausto Giunchiglia, and Bart Selman, editors, *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000)*, pages 717–724, San Francisco, 2000. Morgan Kaufmann.

[23] Raymond Reiter. The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. In Vladimir Lifschitz, editor, *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*, pages 359–380. Academic Press, 1991.

[24] Raymond Reiter. *Knowledge in Action*. MIT Press, Cambridge, Massachusetts, 2001.

[25] Murray Shanahan. *Solving the Frame Problem*. MIT Press, Cambridge, Massachusetts, 1997.

[26] Murray Shanahan. An attempt to formalise a non-trivial benchmark problem in commonsense reasoning. *Artificial Intelligence*, 153(1–2):141–165, 2004.