A First-Order Axiomatization of the Surprise Birthday Present Problem (Preliminary Report): *Abbreviated Version*[*]

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

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 [3], 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 [10]) 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 problems (Yale Shooting Problem [7], Missionaries and Cannibals [10]) but much smaller than large-scale efforts to formalize knowledge, such as the HPKB [14] 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 [12]. First, the goal is to create core, reusable theories of commonsense reasoning, as in [8]. 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 problem, or one invents methods to extend the existing theory. For example, this paper explores how the planning theory of [5] could be extended to joint plans. Third, analyzing a mid-sized problem could result in discovering new representational issues and problems.

Many simplifications are necessary for formalization of even mid-sized problems. The SBP involves time, space, physics, knowledge, perception, naive psychology, multiple agents, and planning. Focusing on all these problems in depth would necessitate a large-scale axiomatization and may lie beyond the capabilities of AI practitioners today.

[*]The full paper can be found at http://www-formal.stanford.edu/leora/sbp.pdf

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

We focus on 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. The formalization of joint plans is an extension of the theory of [5], in which plans consist of a single agent making a request to single or multiple agents, each acting alone.

1.3 Logical Preliminaries

We use a sorted logic. *A, S, T, E, P, Q,* and *X* range, respectively, over agents, situations, calendar-clock-times, events, plans, fluents, and 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 use the situation-based temporal logic of [5]. Time branches forward, but not back. 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*. 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. To do this, we formalize the notion of expectation and extend previous work on the interaction between time, knowledge, and belief. We use the operators *Know* and *Believe* and two corresponding operators describing prediction: *Know-future* and *Expect*.

We note that using only the *Know* and *Know-future* operators would limit the kind of surprise that could be expressed. Consider that one may be surprised by *Q* because one had no expectation that *Q* (*weak surprise*); but one may also, and in a stronger sense, be surprised by because one had the expectation that in fact $\neg Q$ would hold (*strong surprise*). We can use *Know* to express weak but not strong surprise. For it is not possible for *A* to *Know* that $\neg Q$ 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.

We have the usual accessibility relations *K* and *B*, relating, respectively, knowledge-accessible worlds and 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*.

Definition 1 We 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)*

The definitions and axioms that we will have for knowledge and belief are often very similar. We frequently group related definitions and axioms together to save space.

We specify that the *K* relation is reflexive and transitive, and that the *B* relation is symmetric and transitive, yielding an S4 logic of knowledge and a weak S5 logic for belief. This gives the usual axioms on epistemic and doxastic operators [6]. The full paper gives the reason for the choice of semantics.

We place the following restriction on these relations:

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

To see that $B \subseteq K$, note 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.

From Definition 1 and Axiom 1, we have:

Theorem 1 *holds(S, Know(A,Q))* $\Rightarrow$*holds(S, Believe(A,Q))*

To formalize the notion of surprise one must reason about the future. An agent may know (resp. believe) that fluent *Q* will hold at some future time. Thus, we need to reason about the ways in which knowledge (resp. belief) and time interact. An agent may have little or no knowledge about the actions that will be performed. 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.

We assume that the calendar-clock-time structure runs through all possible worlds, and that all agents always know (resp. believe) the calendar-clock-time of the situation they are in.

Axiom 2 *K(A,S1,S2)*$\Rightarrow$*time(S1) = time(S2)*

From Axiom 1, *B(A,S1,S2)*$\Rightarrow$*time(S1) = time(S2)*.

We say that an agent *A* knows (resp. believes) that *Q* will be true at some future time *T* if, for any knowledge (resp. belief) accessible situation *S2*, *Q* will always be true at some situation *S3* later than *S2*, as long as *S3*'s time stamp is *T*.

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

We overload the Know-future/Expect operators so that we can talk about predictions and expectations of event occurrences:

Definition 3 *holds(S1, Know-future(resp. Expect)(A,E, T1))*
$\Leftrightarrow$
$\forall_{S2,S3}$ *K(A,S1,S2) (resp. B(A,S1,S2))* $\wedge$*S2* $<$ *S3* $\wedge$*time(S3) = T1* $\Rightarrow \exists_{S4}$ *occurs(S3,S4,E)*

We now define *A* being surprised at *S1* by a fact *Q* being true or an event *E* occurring starting at *S2*. 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 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.

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

Definition 4 *holds(S1, Surprise(A,Q, S2)) ⇔*
S1 $\geq$ *S2* $\wedge$
holds(S2,Q) $\wedge$
$\forall_{S3<S2}$ *holds(Expect(A,Q,time(S2))* $\Rightarrow$
$\exists_{S4}$ *(S3 < S4 < S2* $\wedge\neg$ *holds(S4, Expect(A,Q,time(S2))))* $\wedge$
$\forall_{S5}$ *K(A,S1,S5)* $\Rightarrow\exists_{S6}$ *S6* $\leq$ *S5* $\wedge$*time(S6) = time(S2)* $\wedge$*holds(S6,Q)* $\wedge$
$\neg\exists_{S7}$ *(S7 < S1* $\wedge\forall_{S5}$ *K(A,S7,S8)* $\Rightarrow\exists_{S9}$ *S9* $\leq$ *S7* $\wedge$*time(S9) = time(S2)* $\wedge$*holds(S9,Q))*

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

Definition 5 *holds(S1, Surprise(A,E, S2)) ⇔*
S1 $\geq$ *S2* $\wedge$
$\exists_{S2*}$ *occurs(S2,S2*,E))* $\wedge$
$\forall_{S3<S2}$ *holds(Expect(A,E,time(S2))* $\Rightarrow$
$\exists_{S4}$ *(S3 < S4 < S2* $\wedge\neg$ *holds(S4, Expect(A,E,time(S2))))* $\wedge$
$\forall_{S5}$ *K(A,S1,S5)* $\Rightarrow\exists_{S6,S6*}$ *S6* $\leq$ *S5* $\wedge$*time(S6) = time(S2)* $\wedge$*occurs(S6,S6*,E)* $\wedge$
$\neg\exists_{S7}$ *(S7 < S1* $\wedge\forall_{S5}$ *K(A,S7,S8)* $\Rightarrow\exists_{S9,S9*}$ *S9* $\leq$ *S7* $\wedge$*time(S9) = time(S2)* $\wedge$*occurs(S9,S9*,E))*

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*. The definition for strong surprise is given in the full paper.

3 Joint plans

Our theory of joint plans extends the theory developed in [5]. 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 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* and *next_step*. *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 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 it 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 [5], 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 Agents acting together

In the theory of [5], agents, even in multiple-agent plans, do not collaborate. For the SBP, we must reason about joint plans in which agents collaborate and act together. We must extend this theory in several ways:
(1) Plan formation. In the original theory, a single requesting agent makes a request of one or more agents. For joint plans, a group of agents jointly decide on a particular plan.
(2) Reserving time blocks. In the original theory, all agents reserve time blocks for all other agents. It is unclear how time blocks will be reserved for joint plans.
(3) Joint actions. The original theory forces asynchronous action: only one agent may act at any particular time. In the SBP, Alice and Bob jointly give Carol her birthday present.

We discuss our approaches to these problems below:

Joint plan formation

We introduce the concet of a joint plan and a joint plan entity (JPE). The JPE 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}. $J \subset A$; this means that all axioms on agents apply to JPEs. An agents is either a joint plan entity (*JPE(A)*) or an individual (*Individual(A)*).

A JPE cannot accept plans from any agent including himself. In fact, no agent is allowed to issue a request to a JPE.

Axiom 3 $\neg\exists_{S1,S2,A,J,P}$ *occurs(S1,S2,request(A,J,P))* $\vee$*accepts_request(P,A,J,S1)*

All joint plans have a similar structure. The JPE 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. The actions in

the JPE may consist of single-agent actions or joint actions, performed by multiple agents. When the JPE's plan succeeds or is abandoned, the JPE ceases to be active.

Axiom 4 *holds(S, active(J)) ⇔*
occurs(S1, S2, broadcast_req(J,members(J),R)) ∧
[S ∈ [S1,S2] ∨
[∃$_A$ A ∈ members(J) ∧assignment(R,A) = P ∧working_on(P,A,J,S2,S)]]

A JPE knows something if all agents in the entity know it:

Axiom 5 *holds(S, Know(J,Q)) ⇔[∀$_A$ (A ∈ members(J) ⇒holds(S, Know(A,Q)))]*

We modify the predicate *governs* which in the original theory ranges over an agent and an action. A JPE's governance should not continue beyond the time that the joint plan is active. We add an extra situational argument to *governs*, and specify that the JPE governs actions only when it is active.

Reserving time blocks: summary
The original theory posited that all agents reserve blocks of time for all agents, including himself. However, one cannot assume that all agents reserve blocks of time for all possible JPEs. 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 reserved blocks of time that *A1* has reserved for *A2* will become reserved for J.

There are several technical points of interest relating to reserving blocks of time for JPEs. These are discussed in detail in the full paper. We summarize the main points below.
(i)We define an *allotment* function that 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 and up to *S*, initially reserved by individual agent *A1* for individual agent *A2*, along with a record of who 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.
(ii) The original theory assumed a maximum delay between successive blocks of time reserved for the same agent. The allotment scheme disturbs this notion, particularly if we allow multiple JPEs created by an identical group of agents: one might never be able to guarantee reserving a block of time or getting anything done (the committee curse). Therefore, we insist that there be no more than one active JPE associated with each group of agents.
(iii) There can still be many JPEs active at any time: $2^n - (n+1)$ non-trivial JPEs for n agents. Agents could become so overcommitted that they cannot successfully execute plans within time constraints. This does not affect our solution to the SBP because n is small. We defer the general problem to futher research.

Joint actions: summary
The main points of this section are summarized below: see the full paper for details.
(i)The extended theory still does not allow concurrency; we merely allow multiple agents to perform a single action.
(ii) It is an axiom of the original theory that agents do not start or end actions at this same time. (This avoids reasoning about the interaction of concurrent actions.) An exception is made at the beginning of time when all agents wait for varying lengths of time. We employ a similar trick for joint actions: agents wait varying amounts of time until they begin the actual performance of the joint action, and then wait varying amounts of time until they begin other actions.
(iii) It is difficult to ensure that the multiple agents involved in a joint action of a JPE will have identical or even overlapping blocks of time reserved for the JPE. We make some assumptions to handle this issue for the SBP, but do not solve the general problem in this paper.

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 first give the plan specification; then discuss the frame problem in this context; and then sketch the proof. This is followed by a paraphrase of the problem premises and domain axioms. The formal statement of the premises and axioms can be found in the full paper.

4.1 Plan Specification

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.

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, the JPE broadcasts a request *r2* to Alice and Bob. At all other times, the JPE waits.

Plan Spec Axiom 1 *next_step(E, p1, S1, S2) ⇔*
action(E, J_{p1}) ∧
first_opportunity(S2, J_{p1}, J_{p1}, S1, p1_f) ⇒
instance(E, broadcast_req(J_{p1}, {alice,bob}, r, S2) ∧
¬first_opportunity(S2,J_{p1},J_{p1}), S1, p1_f) ⇒action(E, J_{p1}) = wait

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

Plan Spec Axiom 2 *holds(S, p1_f) ⇔¬holds(S, in_earshot(carol, bob)) ∧¬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 ∨A = bob ⇒assignment(r,A) = p2*

p1 succeeds if Carol 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.) 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 are taken from [4].

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* $+ \epsilon$*)*

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

A common monotonic solution to the frame problem [11] works by specifying *explanation closure* axioms [15], which state the complete set of actions that can modify a fluent, and by positing *non-occurrence* axioms stating that certain actions do not in fact happen.

We proceed with this approach, rather than using a nonmonotonic solution (e.g. [16]) for two reasons: first, 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 entail. Second, 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. This form of plan specification, therefore, has the potential to reduce the number of non-occurrence assumptions one must make. The connection between plan specification and non-occurrence axioms is a subject for future research.

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, we frequently indicate whether a fact follows from the original theory (O), a lemma in the proof sketch (PS), or the extended theory (ET). [2]

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 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* 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

[2] 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.

will not be informed of the gift giving prior to her birthday. Moreover, Carol will know when she has received her gift.

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

4.4 Domain Axioms

What follows below are English paraphrases; the formal statement is in the full paper on the web.

Premises of the starting situation:

The only individual actors are Alice, Bob, and Carol. In the starting situation Carol does not expect to receive a gift on her birthday. Alice and Bob each have at least $10. The cost of the gift is $20. At the start, neither Alice, Bob, nor Carol owns the gift. Carol is not in earshot of Alice or Bob.

Some housekeeping axioms concerning time: Both Alice and Bob have reserved the entire day of Carol's birthday for themselves. There are two weeks until Carol's birthday. Actions take at most 1/2 hour; *max_delay* is 20 hours.

The joint plan entity J_{p1}, while 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. The governance axioms are very similar to the specification of the forbidden actions in *p2*.

Preconditions on actions:

You can give cash to someone if you own at least that amount of cash. Similarly for earmarking cash for a particular purpose. You can give an object if you physically possess it. You can buy something as long as you have sufficient cash.

Two agents can jointly give an object to a third if:
— one of them physically possesses the object
— both of them have contributed money earmarked toward the object (before the giving of the object). The amount of each agent's contribution must be less than the object's cost; otherwise, the others' earmarking doesn't count.
— both of them have reserved appropriate blocks of time.

Causal axioms:

If one agent gives an object to a second, the first agent no longer has it, and the second does. The transfer of money works similarly. Purchasing an object results in an agent possessing the object but having less money. If someone tells you something, you will believe it.

If *A1* overhears *A3* requesting *A2* to do some plan, he will subsequently know that *A2* has accepted the request to perform that plan.

This axiom 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* may not know that all circumstances crucial for the success of *P* actually hold.

Relations beween actions:

If one has given cash to someone earmarked for some purpose, one has certainly given them cash. Giving entails receiving. One has received a present from someone if there is some person who has given him something.

Knowledge axioms:

In the starting situation, Alice and Bob know all the premises. This means, e.g., that Alice and Bob know at the start that Carol is not in earshot and that the gift costs $20.

Agents always know when it's someone's birthday. Agents always know when they have been involved in a giving, earmarking money, or purchasing action.

Explanation closure axioms:

The only way to have less money is to give it to someone or purchase an item. The only way to lose possession of an item is to give it to someone. The only way to gain possession of an item is to get it from someone or to purchase it.

If one 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. There are only two ways for this to happen: one can overhear such a plan request being issued, or be told that such a plan request has been issued.

Non-occurence axiom: Alice and Bob while J_{p1} is broadcasting the request to do *p2*.

5 Problem Variants

Below, we discuss the variants that the theory can handle, and how we might extend the theory to handle other variants.

Since we have no theory of locations, 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, there will be no JPE, so there is no earmarking of cash, no purchase, no joint gift. As yet, we have not sufficiently formalized the concept of JPE to express or entail what it means when Alice and Bob cannot agree on a joint plan. For similar reasons, we cannot entirely handle the variants where Alice and Bob do not consult together. We can, however, 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 to reason about a variant in which Alice and Bob do not give Carol her gift until after Carol's birthday. One can formulate a plain 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 can give the gift on Carol's birthday. Then although one could show that once Carol gets the gift, she is

surprised, it is possible that it is not on her birthday that Carol is surprised.

The theory can 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. 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 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.

There are two major gaps in the current axiomatization: the lack of 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] is particularly relevant to the first issue. Existing work on negotiation [9] and intentionality [1] 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[13]. Nevertheless, we can tentatively suggest some criteria, as in [12]. 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.

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] P. Cohen and H. Levesque. Intention is choice with commitment. *AIJ*, 42(2–3):263–309, 1990.

[2] E. Davis. Inferring ignorance from the locality of visual perception. In *AAAI1988*, pages 786–790. 1988.

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

[4] E. Davis. A first-order theory of communicating first-order formulas. In *KR2004*, pages 235–245, 2004.

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

[6] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

[7] S. Hanks and D. McDermott. Nonmonotonic logic and temporal projection. *AIJ*, 33(3):379–412, 1987.

[8] P. Hayes. The second naive physics manifesto. In J. and R. Moore, editors, *Formal Theories of the Commonsense World*, pages 1–36. Ablex, Norwood, 1985.

[9] S. Kraus. Negotiation and cooperation in multi-agent environments. *AIJ*, 94(1-2):79–97, 1997.

[10] J. McCarthy. Elaboration tolerance. In *Working Papers, CommonSense 98*, 1998.

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

[12] L. Morgenstern. Mid-sized axiomatizations of commonsense problems. *Studia Logica*, 67(3):353–384, 2001.

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

[14] A. Pease, V. Chaudhri, F. Lehmann, and A. Farquhar. Practical knowledge representation and the DARPA high performance knowledge bases project. In *KR2000*, pages 717–724. Morgan Kaufmann, 2000.

[15] R. Reiter. *Knowledge in Action*. MIT Press, 2001.

[16] M. Shanahan. *Solving the Frame Problem*. MIT Press, 1997.