Available online at www.sciencedirect.com
SCIENCE DIRECT®
Artificial Intelligence 166 (2005) 81–139
Artificial Intelligence
www.elsevier.com/locate/artint

Knowledge and communication: A first-order theory ☆

Ernest Davis

Courant Institute, New York University, New York, NY 10012, USA

Received 1 July 2004; accepted 2 May 2005

Available online 22 June 2005

Abstract

This paper presents a theory of informative communications among agents that allows a speaker to communicate to a hearer truths about the state of the world; the occurrence of events, including other communicative acts; and the knowledge states of any agent—speaker, hearer, or third parties—any of these in the past, present, or future—and any logical combination of these, including formulas with quantifiers. We prove that this theory is consistent, and compatible with a wide range of physical theories. We examine how the theory avoids two potential paradoxes, and discuss how these paradoxes may pose a danger when this theory are extended.

Keywords: Communication; Knowledge; Logic; Paradox

1. Introduction

In constructing a formal theory of communications between agents, the issue of expressivity enters at two different levels: the scope of what can be said *about* the communications, and the scope of what can be said *in* the communications. Other things being equal, it is obviously desirable to make both of these as extensive as possible. Ideally, a theory should allow a speaker to communicate to a hearer truths about the state of the world; the occurrence of events, including other communicative acts; the knowledge states of any

☆ The research reported in this paper was supported in part by NSF grant IIS-0097537.
E-mail address: davise@cs.nyu.edu (E. Davis).

0004-3702/$ – see front matter
doi:10.1016/j.artint.2005.05.002

agent—speaker, hearer, or third parties; any of these in the past, present, or future; and any logical combination of these. This paper presents a theory that achieves pretty much that.

A few examples of what can be expressed in our representation:

(1) Alice tells Bob that all her children are asleep.
(2) Alice tells Bob that she does not know whether he locked the door.
(3) Alice tells Bob that if he finds out who was in the kitchen at midnight, then he will know who killed Colonel Mustard.
(4) Alice tells Bob that no one had ever told her she had a sister.
(5) Alice tells Bob that he has never told her anything she did not already know.

The above examples illustrate many of the expressive features of our representation:

- Example 1 shows that the content of a communication may be a quantified formula.
- Example 2 shows that the content of a communication may refer to knowledge and ignorance of past actions.
- Example 3 shows that the content of a communication may be a complex formula involving both past and future events and states of knowledge.
- Examples 4 and 5 show that the content of a communication may refer to other communications. They also show that the language supports quantification over agents and over the content of a communication, and thus allows the content to be partially characterized, rather than fully specified.

Moreover, our theory supports basic inference from these kinds of representations. For example, given that Alice tells Bob that no one has ever told her that she has a sister, and that Alice knows that, if she did have a sister, someone would have told her, it is possible to infer that Alice knows that she does not have a sister. The proof from our theory of this and similar sample inferences and the representation of the five statements above is given in Section 4.

Following the research programme of [8,21,24,25], the primary purpose of this paper is to develop a representation for expressing commonsense knowledge about knowledge and communication, with the ultimate intention that this representation or something similar, could be used to carry out symbolic reasoning in this domain. A secondary purpose is to develop an object-level theory, expressible in the language, that will justify as broad a range as possible of commonsensically obvious inference in the domain, while entailing as few as possible commonsensically absurd consequences. The success of the language and theory is demonstrated in terms of their ability to capture a large number and a broad range of examples of commonsensically obvious inferences. We are not here concerned with specialized applications, such as distributed systems; with subtle philosophical nuance; or with efficiency of inference in an implemented reasoning engine. Potentially, this theory could find practical application as a logical foundation either for planning communications in a multi-agent system, or for a theory of speech acts to be used in interpreting dialogue or engaging in dialogue with human speakers.

Since our theory allows communications that refer to other communications, and even communications that refer to themselves, there is clearly a danger of running into paradoxes of vicious self-reference. It is therefore particularly important to establish that the theory is consistent. We prove a meta-theorem that the theory is indeed consistent; in fact, that it is consistent with a wide range of domain-specific physical theories and axioms of knowledge acquisition. We discuss two particular apparent paradoxes—an analogue of Russell's paradox, and the "unexpected hanging" paradox—and we show how our theory manages to side-step these.

We should note at the outset the limitations of our theory. The theory deals only with informative acts (and not, for example, with requests) and assumes that the following conditions are true and universally known: If *AS* communicates Q to *AH*, then

(1) *AS* knows that Q is true at the time that he initiates the communication.
(2) From the time that he initiates the communication, *AS* knows that he is carrying out a communication; he knows that the content is Q; and he knows that the recipient is *AH*.
(3) Similarly, when the communication is complete, *AH* knows that he has received a communication; he knows that the content was Q; and he knows that the sender was *AS*.
(4) When the communication is complete, *AS* knows that the communication is complete and *AH* knows the time at which the communication was initiated.

The paradigmatic example of a form of communication satisfying conditions (2), (3), and (4) is direct speech.[1] Another example could be mail, assuming that

- All messages are time-stamped with the time of sending, and signed by the sender.
- There is a universally known maximal delay D between the time of sending and the time of receiving a message. ("Receiving" here means the time when the hearer reads the message, not the time that it arrives in his mailbox.)

In this case, if we define a communication to be "complete" at the time of sending plus D, then the above conditions are met.

Many aspects of the theory can be applied to communications that do not meet condition (4), but I have not been able to find a plausible axiomatization of this more general case that I can prove to be consistent. Also, I cannot prove that the theory is consistent unless time is taken to be discrete. These are discussed further in Section 8.

The paper proceeds as follows: Section 2 reviews the theories of time and of knowledge, which are not new here. Section 3 presents our language and axioms of communication. Section 4 is the core of this paper; it illustrates the power of the theory by showing how it supports the representation of the five sample statements above and three example commonsense inferences. Sections 5 and 6 describe two apparent paradoxes—a paradox analogous to Russell's paradox and the "unexpected hanging" paradox—and explain why these do not cause inconsistencies in the theory. Section 7 gives the statement of Theorems 1 and 2, which assert that the theory is internally consistent and compatible with a wide range of physical theories. Sections 8 and 9 discuss related work. Section 10 dis-

[1] Under assumptions that are reasonable, though not universally valid: e.g. that the speaker knows what he will say when he begins speaking, and that the speaker and hearer have common knowledge that the hearer will correctly understand the message.

cusses open problems and summarizes our conclusions. Appendix A gives the proofs of Theorems 1 and 2.

2. Framework

We use a situation-based, branching theory of time; an interval-based theory of multi-agent actions; and a possible-worlds theory of knowledge. This is all well known, so the description below is brief.

2.1. Time and action

We use a situation-based theory of time. Time can be either continuous[2] or discrete, but it must be *branching*, like the situation calculus. The branching structure is described by the partial ordering "$S1 < S2$", meaning that there is a timeline containing $S1$ and $S2$ and $S1$ precedes $S2$. It is convenient to use the abbreviations "$S1 \leqslant S2$" and "ordered$(S1, S2)$." The predicate "holds(S, Q)" means that fluent Q holds in situation S.

Each agent has, in various situations, a choice about what action to perform next, and the time structure includes a separate branch for each such choice. Thus, the statement that action E is feasible in situation S is expressed by asserting that E occurs from S to $S1$ for some $S1 > S$.

Following McDermott [26], actions are represented as occurring over an interval; the predicate occurs$(E, S1, S2)$ states that action E occurs starting in $S1$ and ending in $S2$. However, the whole theory could be recast without substantial change into the situation calculus extended to permit multiple agents, after the style of Reiter [36]. The advantage of using the "occurs" representation is the much greater ease of extensibility. The situation calculus was developed for domains where a single agent executes a single atomic action in each situation to bring about the next situation; and extending the situation calculus to allow multiple agents, exogenous change, real-valued time, concurrent actions, extended actions, and alternative characterizations of actions involves a series of representational extensions that are somewhat awkward and hard to integrate [36]. By contrast, all of these can be subsumed in the "occurs" representation, though finding a correct axiomatization of a theory with these features can still be difficult.

Table 1 shows the axioms of our temporal theory. Throughout this paper, we use a sorted first-order logic with equality, where the sorts of variables are indicated by their first letter. The sorts are clock-times (T), situations (S), Boolean fluents (Q), actions (E), agents (A), and actionals (Z). An *actional* is a characterization of an action without specifying the agent. For example, the term "puton(blocka, table)" denotes the actional of someone putting block A on the table. The term "do(john, puton(blocka, table))" denotes the action of John putting block A on the table. Free variables in a formula are assumed to be universally quantified.

---

[2] As will be discussed below, I have not proven the theory consistent for continuous theories of time. However, nothing in the form of the representation inherently excludes a continuous model of time; and I conjecture that the theory is, actually, consistent with a continuous model of time.

Table 1
Temporal axioms

Primitives

$T1 < T2$—Time $T1$ is earlier than $T2$.

$S1 < S2$—Situation $S1$ precedes $S2$, on the same time line. (We overload the $<$ symbol.)

time(S)—Function from a situation to its clock time.

holds(S, Q)—Fluent Q holds in situation S.

occurs($E, S1, S2$)—Action E occurs from situation $S1$ to situation $S2$.

do(A, Z)—Function. The action of agent A doing actional Z.

Definitions

TD.1 $S1 \leqslant S2 \equiv S1 < S2 \lor S1 = S2$.

TD.2 ordered($S1, S2$) $\equiv$
$S1 < S2 \lor S1 = S2 \lor S2 < S1$.

TD.3 feasible(E, S) $\Leftrightarrow \exists_{S2}$ occurs($E, S, S2$).

Axioms

T.1 $T1 < T2 \lor T2 < T1 \lor T1 = T2$.

T.2 $\neg[T1 < T2 \land T2 < T1]$.

T.3 $T1 < T2 \land T2 < T3 \Rightarrow T1 < T3$.
(Clock times are linearly ordered.)

T.4 $S1 < S2 \land S2 < S3 \Rightarrow S1 < S3$. (Transitivity)

T.5 $(S1 < S \land S2 < S) \Rightarrow$ ordered($S1, S2$).
(Forward branching)

T.6 $S1 < S2 \Rightarrow$ time($S1$) $<$ time($S2$).
(The ordering on situations is consistent with the orderings of their clock times.)

T.7 $\forall_{S,T1}\exists_{S1}$ ordered($S, S1$) $\land$ time($S1$) $= T1$.
(Every time line contains a situation for every clock time.)

T.8 occurs($E, S1, S2$) $\Rightarrow S1 < S2$.
(Events occur forward in time.)

T.9 [occurs($E, S1, S2$) $\land S1 < SX < S2 \land SX < SY] \Rightarrow$
$\exists_{SZ}\ SX < SZ \land$ ordered(SY, SZ) $\land$ occurs($E, S1, SZ$).
(If action E starts to occur on the time line that includes SY, then it completes on that time line (Fig. 1).)

Note that in our model of time, each feasible action and its consequences are represented by a branch in the time structure. Thus the time structure incorporates everything that *can possibly* happen. We do not single out one particular time line or branch as the history that *will* actually happen. This will be important in our discussion of the paradox of the unexpected hanging.

2.2. Knowledge

As first proposed by Moore [28,29] and widely used since, knowledge is represented by identifying temporal situations with epistemic possible worlds and positing a relation of knowledge accessibility between situations. The relation k_acc(A, S, SA) means that situation SA is accessible from S relative to agent A's knowledge in S; that is, as far as A knows in S, the actual situation could be SA. The statement that A knows ϕ in S is represented by asserting that ϕ holds in every situation that is knowledge accessible from S for A. As is well known, this theory enables the expression of complex interactions of

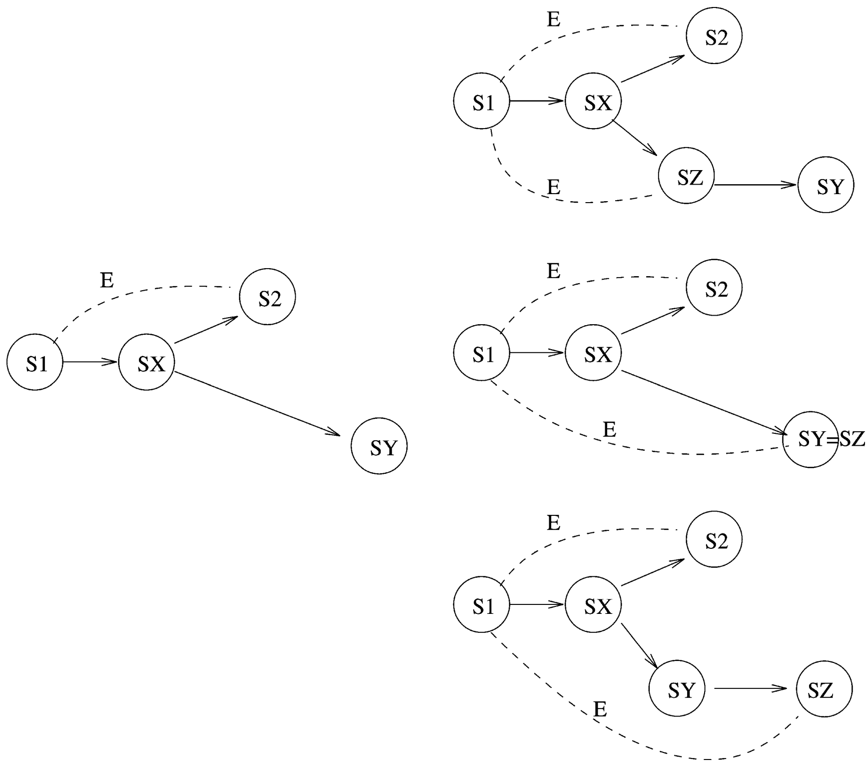

Fig. 1. Axiom T.9. If the time structure has the form on the left, then it has one of the forms on the right.

knowledge and time; one can represent both knowledge about change over time and change of knowledge over time.

Again following Moore [29], the state of agent A knowing *what something is* is expressed by using a quantifier of larger scope than the universal quantification over accessible possible worlds. For example, the statement, "In situation s1, John knows who the President is" is expressed by asserting that there exists a unique individual who is the President in all possible worlds accessible for John from s1.

$$\exists_X \forall_{S1A} \text{k\_acc}(\text{john}, \text{s1}, S1A) \Rightarrow \text{holds}(S1A, \text{president}(X))$$

For convenience, we posit an S5 logic of knowledge; that is, the knowledge accessibility relation, restricted to a single agent, is in fact an equivalence relation on situations. This is expressed in axioms K.1, K.2, and K.3 in Table 2. Three important further axioms govern the relation of time and knowledge.

K.4. Axiom of memory: if A knows ϕ in S, then in any later situation, he remembers that he knew ϕ in S.

K.5. A knows all the actions that he has begun, both those that he has completed and those that are ongoing. That is, he knows a *standard identifier* for these actions; if Bob is dialing (212) 998-3123 on the phone, he knows that he is dialing (212) 998-3123 but

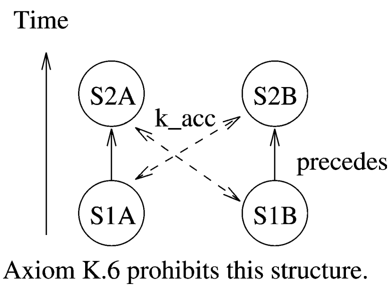

Fig. 2. Axiom K.6.

he may not know that he is calling Ernie Davis. At any time, A knows what actions are feasible for him now.

K.6. Knowledge accessibility relations do not cross in the time structure. (Fig. 2.) In a discrete theory of time, axiom K.6 is a consequence of the axiom of memory K.4. (Knowledge accessibility relations that violate this condition have sometimes been used in the literature for agents who do not satisfy the axiom of memory.)

The theory includes a form of common knowledge, restricted to two agents. Agents $A1$ and $A2$ have *shared knowledge* of ϕ if they both know ϕ, they both know that they both know ϕ and so on.[3] We represent this by defining a further accessibility relation, "sk_acc($A1, A2, S, SA$)" (SA is accessible from S relative to the shared knowledge of $A1$ and $A2$). This is defined as the transitive closure of links of the form k_acc($A1, \cdot, \cdot$) together with links of the form k_acc($A2, \cdot, \cdot$). (Of course, transitive closure cannot be exactly defined in a first-order theory; axioms K.7 and K.8 define an approximation that is adequate for our purposes.)

3. Communication

We now introduce the function "inform", taking two arguments, an agent AH and a fluent Q. The term "inform(AH, Q)" denotes the actional of informing AH that Q; the term "do(AS, inform(AH, Q))" thus denotes the action of speaker AS informing AH that Q. Our theory here treats "do(AS, inform(AH, Q))" as a primitive action; in a richer theory, it would be viewed as an illocutionary description of an underlying locutionary act (not here represented)—the utterance or writing or broadcasting of a physical signal.

We also add a second actional "communicate(AH)". This alternative characterization of a communicative act, which specifies the hearer but not the content of the communication, enables us to separate out *physical* constraints on a communicative act from *contentive* constraints. Thus, we allow a purely physical theory to put constraints on the occurrence

[3] In [10], we need to use common knowledge by a general set of agents. The modifications to the representation and the axioms needed to support this are entirely straightforward.

Table 2
Axioms of knowledge

Primitives

k_acc(A, SA, SB)—SB is accessible from SA relative to A's knowledge in SA.
sk_acc($A1, A2, SA, SB$)—SB is accessible from SA relative to the shared knowledge of $A1$ and $A2$ in SA.

Axioms

K.1 $\forall_{A,SA}$ k_acc(A, SA, SA).

K.2 k_acc(A, SA, SB) $\Rightarrow$ k_ acc(A, SB, SA)

K.3 k_acc(A, SA, SB) $\wedge$ k_ acc(A, SB, SC) $\Rightarrow$ k_acc(A, SA, SC).
(K.1 through K.3 suffice to ensure that the knowledge of each agent obeys an S5 logic: what he knows is true, if he knows ϕ he knows that he knows it; if he does not know ϕ, he knows that he does not know it.)

K.4 [k_acc($A, S2A, S2B$) $\wedge$ $S1A < S2A$] $\Rightarrow$
$\exists_{S1B}$ $S1B < S2B$ $\wedge$ k_acc($A, S1A, S1B$).
(Axiom of memory: If agent A knows ϕ at any time, then at any later time he knows that ϕ was true.)

K.5 [occurs(do(A, Z), $S1A, S2A$) $\wedge$ $S1A \leqslant SA$ $\wedge$
ordered($SA, S2A$) $\wedge$ k_acc(A, SA, SB)] $\Rightarrow$
$\exists_{S1B,S2B}$ occurs(do(A, Z), $S1B, S2B$) $\wedge$
$S1B \leqslant SB$ $\wedge$
[$S2A < SA \Rightarrow S2B < SB$] $\wedge$
[$S2A = SA \Rightarrow S2B = SB$] $\wedge$
[$SA < S2A \Rightarrow SB < S2B$] $\wedge$
[$S1A = SA \Rightarrow S1B = SB$].
(An agent knows which actions he has completed, which actions he has begun, and which actions are now feasible.)

K.6 $\neg\exists_{A,S1A,S1B,S2A,S2B}$
$S1A < S2A$ $\wedge$ $S1B < S2B$ $\wedge$ k_acc($A, S1A, S2B$) $\wedge$ k_acc($A, S2A, S1B$).
(Knowledge accessibility links do not cross in the time structure (Fig. 2).)

K.7 sk_acc($A1, A2, SA, SB$) $\Leftrightarrow$
[k_acc($A1, SA, SB$) $\vee$ k_acc($A2, SA, SB$) $\vee$
sk_acc($A1, A2, SB, SA$) $\vee$
sk_acc($A2, A1, SA, AB$) $\vee$
$\exists_{SC}$ sk_acc($A1, A2, SA, SC$) $\wedge$ sk_acc($A1, A2, SC, SB$)].
(Definition of sk_acc as a equivalence relation, symmetric in $A1, A2$, that includes the k_acc links for the two agents $A1, A2$.)

K.8 (Induction from k_acc links to sk_acc links.) Let $\Phi(S)$ be a formula with a free situational variable S. Then the closure of the following formula is an axiom:
[$\forall_{AS,AH}$[[$\forall_{SA,SB}\Phi(SA)$ $\wedge$ k_acc(AS, SA, SB) $\Rightarrow$ $\Phi(SB)$] $\wedge$
[$\forall_{SA,SB}\Phi(SA)$ $\wedge$ k_ acc(AH, SA, SB) $\Rightarrow$ $\Phi(SB)$]] $\Rightarrow$
[$\forall_{SA,SB}\Phi(SA)$ $\wedge$ sk_ acc(AS, AH, SA, SB) $\Rightarrow$ $\Phi(SB)$].

of a communication, or even to posit physical effects of a communication, but these must be independent of the information content of the communication.

We posit five axioms of communication, summarized in Table 3. Some of these are straightforward; others much less so. We discuss them below in increasing order of complexity. We also put forward a sixth axiom, a frame axiom for ignorance, but its status is more dubious, for reasons that we will discuss.

Table 3
Axioms of communication

I.1	Any inform act is a communication. occurs(do(AS,inform(AH, Q)), $S1$, $S2$) $\Rightarrow$ occurs(do(AS,communicate(AH)), $S1$, $S2$).
I.2	If a speaker AS can communicate with a hearer AH, then AS can inform AH of some specific Q if and only if AS knows that Q holds at the time he begins speaking. feasible(do(AS, communicate(AH)), $S1$) $\Rightarrow$ [$\forall_Q$ feasible(do(AS, inform(AH, Q)), $S1$) $\Leftrightarrow$ [$\forall_{S1A}$ k_acc(AS, $S1$, $S1A$) $\Rightarrow$ holds($S1A$, Q)]]
I.3	If AS informs AH of Q from $S1$ to $S2$, then in $S2$, AH knows that AS has informed him of Q. $\forall_{S1,S2,S2A}$ [occurs(do(AS, inform(AH, Q)), $S1$, $S2$) $\wedge$ k_acc(AH, $S2$, $S2A$)] $\Rightarrow$ $\exists_{S1A}$ occurs(do(AS, inform(AH, Q)), $S1A$, $S2A$) $\wedge$ k_acc(AH, $S1$, $S1A$)
I.4	If AS informs AH of $Q1$ over $[S1, S2]$ and the shared knowledge of AS and AH in $S1$ implies that holds($S1$, $Q1$) $\Leftrightarrow$ holds($S1$, $Q2$), then AS has also informed AH of $Q2$ over $[S1, S2]$. Conversely, the two actions "do(AS,inform(AH, $Q1$))" and "do(AS,inform(AH, $Q2$))" co-occur only if $Q1$ and $Q2$ are related in this way. occurs(do(AS, inform(AH, $Q1$)), $S1$, $S2$) $\Rightarrow$ [occurs(do(AS, inform(AH, $Q2$)), $S1$, $S2$) $\Leftrightarrow$ [$\forall_{S1A}$ sk_acc(AS, AH, $S1$, $S1A$) $\Rightarrow$ [holds($S1A$, $Q1$) $\Leftrightarrow$ holds($S1A$, $Q2$)]]]
I.5	Axiom of comprehension: any property of situations that can be stated in the language is a fluent. Let $\alpha(S)$ be a first-order formula that contains exactly one free variable S of sort "situation" and that does not contain Q as a free variable. (α may have other free variables of other sorts.) Then the closure of the following formula is an axiom: $\exists_Q \forall_S \text{holds}(S, Q) \Leftrightarrow \alpha(S)$. (The closure of a formula β is β scoped by universal quantifications of all its free variables.)
I.6	Frame axiom for ignorance. See the discussion in Section 3.5 below.

3.1. Relation between informing and communication

Axiom I.1. Any inform act is a communication.

$$\text{occurs}(\text{do}(AS, \text{inform}(AH, Q)), S1, S2) \Rightarrow$$
$$\text{occurs}(\text{do}(AS, \text{communicate}(AH)), S1, S2)$$

Axiom I.2. If a speaker AS can communicate with a hearer AH, then AS can inform AH of some specific Q if and only if A knows that Q holds at the time he begins speaking.

$$[\text{feasible}(\text{do}(AS, \text{communicate}(AH)), S1)] \Rightarrow$$
$$[\forall_Q \text{ feasible}(\text{do}(AS, \text{inform}(AH, Q)), S1) \Leftrightarrow$$
$$[\forall_{S1A} \text{ k\_acc}(AS, S1, S1A) \Rightarrow \text{holds}(S1A, Q)]]$$

By virtue of these two axioms, the preconditions for AS informing AH that Q are just that it is feasible for AS to communicate to AH and that AS knows that Q is true. The content Q

may not affect the feasibility in any other way. Axiom I.1 further guarantees that any other physical constraints over communications, such as the duration of a communication or its physical effects, must apply also to inform acts; that is, that the physical characteristics of any inform act must be consistent with the physical constraints on communications. These axioms do not rule out the possibility that the content could affect other physical aspects of the inform act—for example, that a complex content takes longer to communicate than a simple content—but I have not shown that any such constraints lead to a consistent theory.

Note that axiom I.2 requires, conversely, that any fluent Q that is known to be true can be communicated; that is, there is a branch in the time structure corresponding to the communication of Q.

3.2. Epistemic effect of communication

Since we require the strong conditions mentioned in Section 1, we can posit the following axiom:[4]

Axioms I.3. If *AS* informs *AH* of Q from $S1$ to $S2$, then in $S2$, *AH* knows that *AS* has informed him of Q.

$$\forall_{S1,S2,S2A}\ [\text{occurs}(\text{do}(AS, \text{inform}(AH, Q)), S1, S2) \wedge \text{k\_acc}(AH, S2, S2A)] \Rightarrow$$
$$\exists_{S1A}\ \text{occurs}(\text{do}(AS, \text{inform}(AH, Q)), S1A, S2A) \wedge \text{k\_acc}(AH, S1, S1A)$$

Lemmas 3.1 and 3.2 are important consequences of I.3 together with the preceding axioms:

Lemma 3.1. *If AS informs AH of Q, then, when the communication is complete, AS and AH have shared knowledge that the communication has taken place.*

$$\text{occurs}(\text{do}(AS, \text{inform}(AH, Q)), S1, S2) \wedge \text{sk\_acc}(AS, AH, S2, S2A) \Rightarrow$$
$$\exists_{S1A}\ \text{occurs}(\text{do}(AS, \text{inform}(AH, Q)), S1A, S2A)$$

Proof. By K.5, *AS* knows when he has completed a communication.

$$\text{occurs}(\text{do}(AS, \text{inform}(AH, Q)), S1, S2) \wedge \text{k\_acc}(AS, S2, S2A) \Rightarrow$$
$$\exists_{S1A}\ \text{occurs}(\text{do}(AS, \text{inform}(AH, Q)), S1A, S2A)$$

By I.3, *AH* knows when he has received a communication.

$$\text{occurs}(\text{do}(AS, \text{inform}(AH, Q)), S1, S2) \wedge \text{k\_acc}(AH, S2, S2A) \Rightarrow$$
$$\exists_{S1A}\ \text{occurs}(\text{do}(AS, \text{inform}(AH, Q)), S1A, S2A)$$

Choosing $\Phi(S)$ to be the formula "occurs(do(*AS*, inform(*AH*, Q))", the formula in Lemma 3.1 then follows from axiom K.8. □

[4] The statement of this axiom in the KR-2004 paper [9] was not correct.

Lemma 3.2. *If AS informs AH of Q then, when the communication is complete, then AS and AH have shared knowledge that Q was true when the communication began.*

$$\text{occurs}(\text{do}(AS, \text{inform}(AH, Q)), S1, S2) \wedge \text{sk\_acc}(AS, AH, S2, S2A) \Rightarrow$$
$$\exists_{S1A}\ \text{occurs}(\text{do}(AS, \text{inform}(AH, Q)), S1A, S2A) \wedge \text{holds}(S1A, Q)$$

Proof. Let as, ah, q, s0, s1, s2a satisfy the left side of the above implication.
By Lemma 3.1 there exists s1a such that occurs(do(as, inform(ah, q)), s1a, s2a).
By K.1, k_acc(as, s1a, s1a).
By I.2, holds(s1a, q). □

3.3. Axiom of comprehension

The axiom of comprehension states that there is a fluent corresponding to any property of situations definable in the language. The content of this axiom therefore depends on the overall language $\mathcal{L}$. We state this as an axiom schema, i.e., an infinite set of axioms.

Axiom I.5. The comprehension axiom for fluents in a language $\mathcal{L}$ is this: let $\alpha(S)$ be a first-order formula that contains exactly one free variable S of sort "situation" and that does not contain Q as a free variable. (α may have other free variables of other sorts.) Then the closure of the following formula is an axiom:

$$\exists_Q\ \forall_0 S\ \text{holds}(S, Q) \Leftrightarrow \alpha(S)$$

(The closure of a formula β is β scoped by universal quantifications of all its free variables.)

What this axiom states is that, given any property $\alpha(S)$, there exists a fluent Q that holds on just those situation satisfying α. Moreover, the property α may be parameterized by free variables. The form of this axiom is modelled on the formulation of the "separation" or "subset" axiom of ZF set theory as given in, for example, [15], which states that, given any set B and property α, there exists a subset $C \subset B$ of all the elements in B satisfying α. In axiom I.5, the set of all situations corresponds to B and the fluent Q corresponds to the subset C. Further discussion is given in Appendix A, particularly in Lemma A.21.

Let us first discuss the significance of free variables in the formula α. The reason to allow free variables that are not situations is to deal with examples such as the following: We want to be able to posit that a speaker can say, for example, that some specific block is either red or blue without requiring that the language $\mathcal{L}$ have a constant symbol for each block, or even a formula that uniquely identifies each block.[5]

[5] You might well ask, "If you cannot refer to the block in $\mathcal{L}$, how is the speaker talking about it?" Perhaps he is pointing. Perhaps he is using a richer language with more constant symbols. The language $\mathcal{L}$ does not have to be the language that the speaker is actually using; it is a language in which we, externally, describe what the speaker is saying. It is not a very important point, but it does make the theory more elegant and easier to use if one assumes that a speaker can refer *de re* to any entity other than a situation.

This axiom achieves this. We choose $\alpha(S)$ to be the formula "holds(S, red(X)) $\vee$ holds(S, blue(X))". The axiom schema then state

$$\forall_X \exists_Q \forall_S \text{holds}(S, Q) \Leftrightarrow \text{holds}(S, \text{red}(X)) \vee \text{holds}(S, \text{blue}(X))$$

That is, for every object X there is a fluent Q that corresponds to the situations in which X is either red or blue.

The reason to exclude formulas that have other situational free variables in addition to S is that it does not seem to mean anything to have this kind of *de re* reference to situations. A situation is meaningful only in relation to the current situation; there is no other way to meaningfully refer to a situation. It may be noted that the consistency proof for the theory (Theorem 1 below) does not depend on this restriction.

The reason for the condition that Q not appear free in the formula is that, otherwise, we could choose $\alpha(S)$ to be the formula $\neg$holds(S, Q), in which case the axiom would give us $\exists_Q \forall_S$ holds(S, Q) $\Leftrightarrow$ $\neg$holds(S, Q), which is obviously not satisfiable.[6] Note, however, that if a different variable name is chosen, there is no problem with having a free variable of sort "fluent". For example, if we choose $\alpha(S)$ to be the formula $\neg$holds(S, $Q1$), then the schema yields the axiom $\forall_{Q1} \exists_Q \forall_S$ holds(S, Q) $\Leftrightarrow$ $\neg$holds(S, $Q1$) which is entirely reasonable.

The content of the comprehension axiom depends on the overall language $\mathcal{L}$. In general, one supposes that the language $\mathcal{L}$ will contain many domain and problem specific symbols beyond those that are used in the axioms enumerated here. Theorem 1 shows that these axioms are consistent when $\mathcal{L}$ is a physical language augmented with the symbols from the theory of knowledge and communication described here. In [10] we consider a language that includes also agent commitments and requests. In that setting, the above formulation of the axiom turns out to be too strong; we have to limit the comprehension axiom to apply only to formulas that do not include symbols describing commitment and requests.

In view of this comprehension axiom, axiom K.8 could be restated as a single axiom (rather than an axiom schema) as follows:

K.8.A. $\forall_{Q,AS,AH}$ [[$\forall_{S,SA}$ holds(S, Q) $\wedge$ k_acc(AS, S, SA) $\Rightarrow$ holds(SA, Q)] $\wedge$
[$\forall_{S,SA}$ holds(S, Q) $\wedge$ k_acc(AH, S, SA) $\Rightarrow$ holds(SA, Q)]] $\Rightarrow$
[$\forall_{S,SA}$ holds(S, Q) $\wedge$ sk_acc(AS, AH, S, SA) $\Rightarrow$ holds(SA, Q)].

However we did not use this formulation originally because we did not want K.8 to be dependent on I.5.

3.4. Independent actions

In a temporal representation, like ours, that permits the concurrent execution of actions, it does not suffice just to describe what actions can be executed; one must also, to greater or lesser extent, describe what combinations of actions can be executed concurrently. At the minimum, if two actions are independent, it should be possible to execute the one without

[6] I am extremely grateful to the anonymous reviewer who pointed this out.

the other. In the case of "inform" acts, the natural axiom would be that, if AS knows ϕ, then he can choose to carry out the single act of informing AH of ϕ and not doing anything else. One might suppose that this could be expressed in the following two axioms:

WRONG.1 $\text{feasible}(\text{do}(AS, \text{inform}(AH, Q)), S1) \Rightarrow$
$\exists_{S2}\ \text{occurs}(\text{do}(AS, Z), S1, S2) \Leftrightarrow Z = \text{do}(AS, \text{inform}(AH, Q))$.

WRONG.2 $\text{do}(AS1, \text{inform}(AH1, Q1)) = \text{do}(AS2, \text{inform}(AH2, Q2)) \Rightarrow$
$AS1 = AS2 \wedge AH1 = AH2 \wedge Q1 = Q2$.

However, as my labels subtly suggest,[7] this is not an acceptable formulation. In fact, as we shall show in Section 5, these are inconsistent with the axiom of comprehension I.5.

The problem, intuitively, is this: The comprehension axiom asserts that there exists a fluent for every set of situations; axiom WRONG.1 asserts that there exists a separate branch in time for every fluent. Therefore, if you try to construct a model of these axioms combined, you first have to construct all sets of situations; then add branches for each of these, which gives a whole bunch more resultant situations; these in turn generate vast numbers of new sets of situations .... There is no way to make this construction converge. (I'm being a little loose here, but one can make this tight. The decisive proof that this cannot be made to work is the "misled" paradox of Section 5.)

Therefore, we have to weaken axiom WRONG.1.[8] The approach we take is as follows: In general, it is only necessary to distinguish an occurrence of action A1 from an occurrence of action A2 if they have different causal consequences. For instance, in the blocks world, if all you are interested in is the sequence of towers that are formed, then all that matters in discriminating actions is the ending position of the block being moved; the trajectory through which it moves is immaterial.

Now, in the case of informative acts, the causal consequence of concern is the effect on knowledge states. Assuming axiom I.3, the main effect of AS informing AH of Q is that, when the communication is complete, AS and AH have shared knowledge that Q held at the beginning of the communication. Therefore, if $Q1$ and $Q2$ are two informative contents such that the effects on the shared knowledge of AS and AH following a communication of $Q1$ from AS to AH are the same as those effects following a communication of $Q2$, then we can treat the communication of $Q1$ and the communication of $Q2$ as *the same action*; they, so to speak, attain the same end state via different trajectories. And a sufficient condition to ensure this is that AS and AH have shared knowledge at the *start* of the communication that $Q1$ and $Q2$ are equivalent.

For example, if Jack and Jane share the knowledge that George Bush is the President and that 1600 Pennsylvania Avenue is the White House, then the action of Jack informing

[7] One thing I have learned in twenty years of teaching is that, if you write down a wrong formula on the blackboard for purposes of discussion, you have to label it WRONG in large letters. Otherwise, students copy it into their notebooks .... Similarly, if someone is skimming through this paper looking for formal axioms, I do not want him to use these.

[8] Weakening axiom WRONG.2 does not work. In fact, WRONG.2 ends up being true in the model we will construct, but its truth won't actually matter much once we have correctly reformulated WRONG.1.

Jane that Bush is at the White House is identical to the act of Jack informing Jane that the President is at 1600 Pennsylvania Avenue. If they do not share this knowledge, then these two acts are different.

This, then, is our axiom: The event of AS informing AH of $Q1$ and the event of AS informing AH of $Q2$ co-occur over an interval $[S1, S2]$ if and only if AS and AH have shared knowledge in $S1$ that $Q2$ if and only if $Q1$,

I.4. $\text{occurs}(\text{do}(AS, \text{inform}(AH, Q1)), S1, S2) \Rightarrow$
$[\text{occurs}(\text{do}(AS, \text{inform}(AH, Q2)), S1, S2) \Leftrightarrow$
$[\forall_{S1A}\ \text{sk\_acc}(AS, AH, S1, S1A) \Rightarrow$
$[\text{holds}(S1A, Q1) \Leftrightarrow \text{holds}(S1A, Q2)]]]$.

As we shall see in Section 5, in a discrete model of time this is sufficient to avoid the contradiction.

Note: The above axiom is not sufficient to rule out models in which the informative actions of one agent constrain the concurrent actions of another agent. The easiest way to insure independence between agents is to posit an axiom of "anti-synchrony" that no two agents begin two actions at the same time [36].

T.10. $\text{occurs}(\text{do}(A1, Z1), S1, S2) \wedge \text{occurs}(\text{do}(A2, Z2), S1, S3) \Rightarrow A1 = A2$.

However, since this axiom is part of the physical theory, and not all physical theories may wish to use it, we have not made it part of our standard set of temporal axioms.

Two alternative formulations of axiom I.4 should be mentioned. We can weaken I.4 to read that communicating $Q1$ and $Q2$ co-occur just if they coincide over all situations of the same time as the beginning of the situation.

I.4.A. $\text{occurs}(\text{do}(AS, \text{inform}(AH, Q1)), S1, S2) \Rightarrow$
$[\text{occurs}(\text{do}(AS, \text{inform}(AH, Q2)), S1, S2) \Leftrightarrow$
$[\forall_{S1A}\ \text{time}(S1A) = \text{time}(S1) \Rightarrow$
$[\text{holds}(S1A, Q1) \Leftrightarrow \text{holds}(S1A, Q2)]]]$.

The consistency proof in Appendix A requires only a small modification to deal with this new version. However, this version seems to me harder to justify than the previous version.

A second alternative, which is in effect equivalent to axiom I.4.A, is to use the axioms WRONG.1 and WRONG.2 and modify the comprehension axiom to state that there is a fluent corresponding to every property of situations at some particular time T:

I.5.B. Let $\alpha(S)$ be a first-order formula in $\mathcal{L}$ with exactly one free variable S of sort "situation", in which the variable Q does not appear free. (α may have other free variables of other sorts.) Then the closure of the following formula is an axiom:

$$\forall_T\ \exists_Q\ \forall_S\ \text{holds}(S, Q) \Leftrightarrow \alpha(S) \wedge \text{time}(S) = T$$

3.5. The frame inference

Finally, it would be desirable to carry out the frame inference over knowledge and ignorance.

The frame axiom over knowledge is just the axiom of memory, axiom K.4: if A knows in S that ϕ is true, then he remembers in all later situations that ϕ was true. Since we have no actions or events that cause forgetting, this simple formulation suffices. Note that "knowing ϕ" is represented as "all worlds in which ϕ is false are inaccessible." Hence preserving knowledge amounts to saying that if situation *SB* is inaccessible from *SA* then any temporal descendant of *SB* is inaccessible from the corresponding descendant of *SA*.

The frame axiom over ignorance is the reverse: Given that A does not know ϕ in $S0$, and given that nothing occurs between $S0$ and $S1$ that would cause him to learn ϕ, we wish to infer that he still does not know ϕ in $S1$. Since "not knowing ϕ in S" is represented as "there are possible worlds accessible from S in which ϕ is false," this frame inference should have the following general form: If $S0A$ is accessible from $S0$, $S1 > S0$, $S1A > S0A$, and as far as A's sources of knowledge are concerned, the interval between $S0$ and $S1$ is indistinguishable from the interval between $S0A$ and $S1A$, then $S1A$ is accessible from $S1$.

Stating this formally is mostly a matter of collecting all the necessary sources of knowledge. Our theory requires that agent A gains knowledge in S under the following circumstances

(1) If A begins action E in $S1$, and $S2$ is on a branch in which E is executed, then in $S2$, A knows that E is executed. If E is completed at or before $S2$, then in $S2$ A knows when it was completed.
(2) If action E is feasible for A in situation S, then A knows that E is feasible in S.
(3) If A receives a communication from *AS* in S then A knows in S that he has received a communication.

We also assume that there are domain-specific axioms of knowledge production. In an S5 logic, it is reasonable to assume that these are all of the following form: In all situations S, A knows whether $\Phi(A, S)$, where Φ is a formula that can refer only to present or past *physical* states or to past (but not present) knowledge states.[9] Formally, we impose the following conditions on $\Phi(A, S)$:

- The only free variables in $\Phi(A, S)$ are A and S.
- If $S1$ is a quantified variable other than S appearing in Φ, and $S1$ is used as either the second-to-last or last argument for either k_acc or sk_acc, then the quantification of $S1$ imposes the restriction $S1 < S$.
- If $S1$ is a quantified variable other than S appearing in Φ, and $S1$ is not used as an argument for either k_acc or sk_acc, then the quantification of $S1$ imposes the restriction $S1 \leqslant S$.

[9] Actually, I conjecture that these restrictions are not necessary, and that it is consistent to allow Φ to be any formula, but I have not proven it.

Table 4
Frame action for ignorance

I.6: $[\text{k\_acc}(A, S0A, S0B) \wedge S0A < S1A \wedge S0B < S1B \wedge \text{time}(S1B) = \text{time}(S0B) \wedge$

$$
\begin{aligned}
(1)\quad & [\forall_{S2A,S3A,Z}\ [\text{occurs}(\text{do}(A, Z), S2A, S3A) \wedge S2A \leqslant S1A \wedge S0A < S3A \wedge \\
& \quad \text{ordered}(S1A, S3A)] \Rightarrow \\
& \quad \exists_{S2B,S3B}\ \text{occurs}(\text{do}(A, Z), S2B, S3B) \wedge \text{time}(S2B) = \text{time}(S2A) \wedge \\
& \qquad [[S1A < S3A \wedge S1B < S3B] \vee \\
& \qquad [S3B \leqslant S1B \wedge \text{time}(S3B) = \text{time}(S3A)]]] \wedge \\
(2)\quad & [\forall_{S2B,S3B,Z}\ [\text{occurs}(\text{do}(A, Z), S2B, S3B) \wedge S2B \leqslant S1B \wedge S0B < S3B \wedge \\
& \quad \text{ordered}(S1B, S3B)] \Rightarrow \\
& \quad \exists_{S2A,S3A}\ \text{occurs}(\text{do}(A, Z), S2A, S3A) \wedge \text{time}(S2A) = \text{time}(S2B) \wedge \\
& \qquad [[S1B < S3B \wedge S1A < S3A] \vee \\
& \qquad [S3A \leqslant S1A \wedge \text{time}(S3A) = \text{time}(S3B)]]] \wedge \\
(3)\quad & [\forall_{S2A,S3A,AS,Q}\ [\text{occurs}(\text{do}(AS, \text{inform}(A, Q)), S2A, S3A) \wedge S3A \leqslant S1A] \Rightarrow \\
& \quad \exists_{S2B,S3B}\ \text{occurs}(\text{do}(AS, \text{inform}(A, Q)), S2B, S3B) \wedge S3B \leqslant S1B \wedge \\
& \qquad \text{time}(S2B) = \text{time}(S2A) \wedge \text{time}(S3B) = \text{time}(S3A)] \wedge \\
(4)\quad & [\forall_{S2B,S3B,AS,Q}\ [\text{occurs}(\text{do}(AS, \text{inform}(A, Q)), S2B, S3B) \wedge S3B \leqslant S1B] \Rightarrow \\
& \quad \exists_{S2A,S3A}\ \text{occurs}(\text{do}(AS, \text{inform}(A, Q)), S2A, S3A) \wedge S3A \leqslant S1A \wedge \\
& \qquad \text{time}(S2A) = \text{time}(S2B) \wedge \text{time}(S3A) = \text{time}(S3A)] \wedge \\
(5)\quad & [\forall_{S2A,S2B}\ [S2A \leqslant S1A \wedge S2B \leqslant S1B \wedge \text{time}(S2A) = \text{time}(S2B)] \Rightarrow \\
& \quad \bigwedge_i [\Phi_i(S2A) \Leftrightarrow \Phi_i(S2B)]] \\
]& \\
\Rightarrow & \ \text{k\_acc}(A, S1A, S1B).
\end{aligned}
$$

Thus we assume the existence of a finite collection of axioms of the form

$$
\begin{aligned}
&\forall_{A,S}\ [[\forall_{SA}\text{k\_acc}(A, S, SA) \Rightarrow \Phi_i(A, S)] \vee \\
&\quad [\forall_{SA}\ \text{k\_acc}(A, S, SA) \Rightarrow \neg\Phi_i(A, S)]]
\end{aligned}
$$

For example, Scherl and Levesque [38,39] propose the use of an action "SENSE$_Q$" which informs the actor whether fluent Q is true. We can achieve that in the above framework by choosing $\Phi(A, S)$ to be the condition that A has executed SENSE$_Q$ and Q holds:

$$\Phi(A, S) \Leftrightarrow \exists_{S1}\ \text{occurs}(\text{SENSE}_Q, S1, S) \wedge \text{holds}(S, Q)$$

We now posit that every agent always knows whether $\Phi(A, S)$. Since, by axiom K.5, an agent always knows whether he has executed SENSE$_Q$, it follows that, if an agent has executed SENSE$_Q$, then he knows whether Q is true.

So now we can state the frame axiom I.6 asserting that if none of the above conditions has been met, then a knowledge accessibility relation persists. (Table 4.)

That is: suppose that $S0B$ is knowledge accessible from $S0A$ relative to A, $S1A$ follows $S0A$, $S1B$ follows $S1A$, $S1A$ and $S1B$ have the same clock-time, and the following conditions are met:

(1) If A executes actional Z, either completing it or starting it between $S1A$ or beginning it at $S2A$, then he executes the same action at the corresponding times in the interval $[S0B, S1B]$. (If the action ends after $S1A$ and $S1B$, then the clock-times of the endings need not be the same.)
(2) The reverse of 1; if A executes an action in the "B" interval then he executes the same action at the corresponding time in the "A" interval.
(3) If *AS* tells A of Q and completes this action between $S0A$ and $S1A$, then the same thing happens between $S0B$ and $S1B$.
(4) The reverse of (3): If *AS* tells A of Q and completes this action between $S0B$ and $S1B$, then the same thing happens between $S0A$ and $S1A$.
(5) All of the facts Φ_i have the same truth value from $S0A$ to $S1A$.

Then nothing that A knows about has occurred to distinguish the interval $[S0B, S1B]$ from the interval $[S0A, S1A]$, and therefore $S1B$ is knowledge accessible from $S1A$.

Well, there it is. It is not a candidate for any "Top 10 most elegant axioms" lists.

A more serious problem is that it does not give us what we want. What we want is: Given that in s0, Sam does not know whether Herbert Hoover invented the vacuum cleaner (P), and given that the only thing that happens between s0 and s1 is that Jack tells Sam that tea is selling for \$2 a pound in Shanghai (Q), we should be able to infer that Jack still does not know whether Herbert Hoover invented the vacuum cleaner. But that inference is not valid. The problem is that it is consistent with the givens that Sam originally knows $\neg P \Leftrightarrow Q$, and so, when Jack tells him Q he finds out $\neg P$. Alternatively, Sam may originally know the weaker statement, "If Jack knows Q, then P;" again, when Jack tells him Q he can infer that Jack knows Q and therefore P.

The problem here is not with the frame axiom; the frame axiom is fine. The problem is with the specification of the initial state. You need to add the condition that the agent does not know anything except the givens. Halpern [17] presents a multi-agent model in which an agent knows only a specific collection of statements and their logical consequences, and nothing more about the world including other agents' knowledge (more precisely, he presents a collection of such theories corresponding to different models of knowledge); and similar studies have been done by Levesque [22]. The problem, though, is that these theories only work in the case where we can specify everything that the agent knows. In most real cases, we do not know everything that Sam knows, but we still want to make the inference. How this inference can be characterized is entirely an open question; and once it is solved (perhaps non-monotonically) it is unclear whether it would use axiom I.6 at all. It would be hard to find any plausible commonsense inference problems where axiom I.6 was useful. (In [7], I have studied a special case of this frame inference where the occurrence of an event is physicially hidden from an agent, and it is therefore possible to infer that the agent remains ignorant of it.)

4. Sample inferences

We now illustrate the power of the above theory by showing how the sample scenarios in the introduction can be represented, and how three toy inferences can be justified.

Table 5
Notational extensions

	Definitions
KD.1	holds(S, know(A, $Q1$)) $\equiv$ [$\forall_{SA}$ k_acc(A, S, SA) $\Rightarrow$ holds(SA, $Q1$)].
KD.2	holds(S, not($Q1$)) $\equiv$ $\neg$holds(S, $Q1$).
KD.3	holds(S, know_whether(A, $Q1$)) $\equiv$ holds(S, know(A, $Q1$)) $\vee$ holds(S, know(A, not($Q1$)))
KD.4	Let $\beta(S)$ be a formula with a free variable S (and possibly other free variables). Let μ be a variable of sort "fluent" that does not appear free in β and let $\Phi(\mu)$ be a formula. The expression "$\Phi(\lambda(S)\beta(S))$" should be expanded to read $\exists_\mu\, [\forall_S \text{ holds}(S,\mu) \Leftrightarrow \beta(S)] \wedge \Phi(\mu)$ In an expression with multiple lambda expressions, the expressions should be expanded from left to right, from outside to inside.

To help make the representations more readable and more elegant, we will begin by defining four further notations (Table 5). First we define "know(A, Q)" as a function mapping agent A and fluent Q to the fluent of A knowing that Q holds in S; that is, Q holds in all situations accessible from S (definition KD.1).

The existence of such a fluent is guaranteed by the comprehension axiom. Let $\alpha(S)$ be the open formula "$\forall_{SA}$ k_acc(A, S, SA) $\Rightarrow$ holds(SA, $Q1$)". Then the comprehension schema asserts

$$\forall_{A,Q1}\, \exists_Q\, \forall_S \text{ holds}(S, Q) \Leftrightarrow \forall_{SA} \text{ k\_acc}(A, S, SA) \Rightarrow \text{holds}(SA, Q1)$$

For any particular A and $Q1$, the fluent Q satisfying this property has the property we need for know(A, $Q1$). Note that, in this construal "know" is a garden-variety first-order function both in its syntax and its semantics.

Second, we define "not(Q)" as the function mapping fluent Q to the fluent of Q not holding. Third, we define "know_whether(A, Q)" as a function mapping agent A and fluent Q to the fluent of A knowing whether or not Q is true. Again, the existence of such fluents is guaranteed by the comprehension axiom, and these are simple first-order functions.

The final notation is a macro extension to first-order syntax ("syntactic sugar"). We will use expressions of the form $\lambda(S)\beta(S)$ to denote the fluent that holds in situation S just if formula β holds of S. Thus, for examples, the fluent that Joe has just completed putting block A on B can be denoted by the expression

$$\lambda(S)\exists_{S0} \text{ occurs}(\text{do}(\text{joe}, \text{puton}(a, b)), S0, S)$$

The statement that Sam knows in situation s1 that Joe has just completed putting block A onto B can thus be expressed

$$\text{holds}(\text{s1}, \text{know}(\text{sam}, \lambda(S)\exists_{S0} \text{ occurs}(\text{do}(\text{joe}, \text{puton}(a, b)), S0, S)))$$

These lambda expressions are defined within our theory as macros that expand into first-order formulas. (It should be emphasized that we are not here defining a *general* lambda calculus, just lambda expressions with one situational argument and a fluent value.) The expansion rule is given in definition KD.4 in Table 5.

For example, the formula

$$\text{holds}(\text{s1}, \text{know}(\text{sam}, \lambda(S)\exists_{S0} \text{ occurs}(\text{do}(\text{joe}, \text{puton}(a, b)), S0, S)))$$

expands to the formula

$$\exists_Q\ [\forall_S\ \text{holds}(S, Q) \Leftrightarrow \exists_{S0}\ \text{occurs}(\text{do}(\text{joe}, \text{puton}(a, b)), S0, S)] \wedge$$
$$\text{holds}(\text{s1}, \text{know}(\text{sam}, Q))$$

Using the definition of "know", this is equivalent to

$$\exists_Q\ [\forall_S\ \text{holds}(S, Q) \Leftrightarrow \exists_{S0}\ \text{occurs}(\text{do}(\text{joe}, \text{puton}(a, b)), S0, S)] \wedge$$
$$\forall_{S1A}\ \text{k\_acc}(\text{sam}, \text{s1}, S1A) \Rightarrow \text{holds}(S1A, Q)$$

Since the existence of a fluent Q satisfying this first line is guaranteed by the comprehension axiom, this is equivalent to

$$\forall_{S1A}\ \text{k\_acc}(\text{sam}, \text{s1}, S1A) \Rightarrow$$
$$\exists_{S0}\ \text{occurs}(\text{do}(\text{joe}, \text{puton}(a, b)), S0, S1A).$$

In the examples that follow, we will give both the compacted representation (with "know" and lambda expressions) and the expanded versions without them.

4.1. *Sample representations*

We illustrate the expressive power of our representation using the examples from the introduction.

4.1.1. *Sample representation 1*

Alice tells Bob that all her children are asleep.

$$\text{occurs}(\text{do}(\text{alice}, \text{inform}(\text{bob},$$
$$\lambda(S)\ \forall_C\ \text{holds}(S, \text{child}(C, \text{alice})) \ \Rightarrow \text{holds}(S, \text{asleep}(C)))),$$
$$\text{s0}, \text{s1})$$

In expanded form:

$$\exists_Q\ \text{occurs}(\text{do}(\text{alice}, \text{inform}(\text{bob}, Q)), \text{s0}, \text{s1}) \wedge$$
$$\forall_S\ \text{holds}(S, Q) \Leftrightarrow$$
$$[\forall_C\ \text{holds}(S, \text{child}(C, \text{alice})) \ \Rightarrow \text{holds}(S, \text{asleep}(C))]$$

4.1.2. *Sample representation 2*

Alice tells Bob that she does not know whether he locked the door.

$$\text{occurs}(\text{do}(\text{alice}, \text{inform}(\text{bob},$$
$$\lambda(S)\ \text{holds}(S, \text{not}(\text{know\_whether}(\text{alice},$$
$$\lambda(SA)\ \exists_{S1A,S2A}\ \text{occurs}(\text{do}(\text{bob}, \text{lock\_door}), S1A, S2A) \wedge S1A < SA$$
$$))))),$$
$$\text{s0}, \text{s1})$$

Expanding and rearranging gives

$$
\begin{array}{l}
\exists_Q \text{ occurs(do(alice, inform(bob, } Q)), \text{s0, s1}) \wedge \\
\quad \forall_S \text{ holds}(S, Q) \Leftrightarrow \\
\quad\quad [\exists_{SA} \text{ k\_acc(alice, } S, SA) \wedge \\
\quad\quad\quad \exists_{S1A,S2A}\ S1A < S2A < SA \wedge \\
\quad\quad\quad\quad \text{occurs(do(bob, lock\_door)}, S1A, S2A)] \wedge \\
\quad\quad [\exists_{SA} \text{ k\_acc(alice, } S, SA) \wedge \\
\quad\quad\quad \neg\exists_{S1A,S2A}\ S1A < S2A < SA \wedge \\
\quad\quad\quad\quad \text{occurs(do(bob, lock\_door)}, S1A, S2A)]
\end{array}
$$

4.1.3. Sample representation 3

Alice tells Bob that if he finds out who was in the kitchen at midnight, then he will know who killed Colonel Mustard. (Note: The interpretation below assumes that exactly one person was in the kitchen at midnight.)

$$
\begin{array}{l}
\text{occurs(do(alice, inform(bob,} \\
\quad \lambda(S)\ \forall_{SA}\ [S < SA \wedge \\
\quad\quad \exists_{PK} \text{ holds}(SA, \text{know(bob,} \\
\quad\quad\quad \lambda(SC)\ \exists_{S3C}\ S3C < SC \wedge \text{time}(S3C) = \text{midnight} \wedge \\
\quad\quad\quad\quad \text{holds}(S3C, \text{in}(PK, \text{kitchen}))))] \\
\quad\quad \Rightarrow \\
\quad\quad \exists_{PM} \text{ holds}(SA, \text{know(bob,} \\
\quad\quad\quad \lambda(SB)\ \exists_{S2B,S3B}\ S3B < SB \wedge \\
\quad\quad\quad\quad \text{occurs(do}(PM, \text{kill(mustard))}, S2B, S3B))))), \\
\text{s0, s1})
\end{array}
$$

Expanding and rearranging gives:

$$
\begin{array}{l}
\exists_Q \text{ occurs(do(alice, inform(bob, } Q)), s0, s1) \wedge \\
\quad \forall_S \text{ holds}(S, Q) \Leftrightarrow \\
\quad\quad \forall_{S2}\ [S2 > S \wedge \\
\quad\quad\quad \exists_{PK} \forall_{S2A} \text{ k\_acc(bob, } S2, S2A) \Rightarrow \\
\quad\quad\quad\quad \exists_{S3A}\ S3A < S2A \wedge \text{midnight(time}(S3A)) \wedge \\
\quad\quad\quad\quad \text{holds}(S3A, \text{in}(PK, \text{kitchen}))] \Rightarrow \\
\quad\quad\quad [\exists_{PM} \forall_{S2B} \text{ k\_acc(bob, } S2, S2B) \Rightarrow \\
\quad\quad\quad\quad \exists_{S3B,S4B}\ S3B < S4B < S2B \wedge \\
\quad\quad\quad\quad\quad \text{occurs(do}(PM, \text{murder(mustard))}, S3B, S4B)]
\end{array}
$$

4.1.4. Sample representation 4

Alice tells Bob that no one had ever told her she had a sister.

$$\begin{array}{l}
\text{occurs}(\text{do}(\text{alice}, \text{inform}(\text{bob}, \\
\quad\quad \lambda(S)\ \neg\exists_{AP,S1,S2}\ S2 < S \wedge \\
\quad\quad\quad \text{occurs}(\text{do}(AP, \text{inform}(\text{alice}, \\
\quad\quad\quad\quad \lambda(SA)\ \exists_{P2}\ \text{holds}(SA, \text{sister}(P2, \text{alice})))) \\
\quad\quad\quad\quad S1, S2))) \\
\quad \text{s0}, \text{s1})
\end{array}$$

Expanding and rearranging,

$$\begin{array}{l}
\exists_Q\ \text{occurs}(\text{do}(\text{alice}, \text{inform}(\text{bob}, Q)), \text{s0}, \text{s1}) \wedge \\
\quad \forall_S\ \text{holds}(S, Q) \Leftrightarrow \\
\quad\quad \neg\exists_{S2,S3,Q1,P1}\ S2 < S3 < S \wedge \\
\quad\quad\quad \text{occurs}(\text{do}(P1, \text{inform}(\text{alice}, Q1)), S2, S3) \wedge \\
\quad\quad\quad \forall_{SX}\ \text{holds}(SX, Q1) \Rightarrow \exists_{P2}\ \text{holds}(SX, \text{sister}(P2, \text{alice}))
\end{array}$$

4.1.5. Sample representation 5

Alice tells Bob that he has never told her anything she didn't already know.

$$\begin{array}{l}
\text{occurs}(\text{do}(\text{alice}, \text{inform}(\text{bob}, \\
\quad\quad \lambda(S)\ \forall_{S2,S3,Q}\ S3 \leqslant S \wedge \text{occurs}(S2, S3, \text{do}(\text{bob}, \text{inform}(\text{alice}, Q))) \Rightarrow \\
\quad\quad\quad \text{holds}(S2, \text{know}(\text{alice}, Q)))) \\
\quad \text{s0}, \text{s1})
\end{array}$$

Expanding and rearranging gives:

$$\begin{array}{l}
\exists_Q\ \text{occurs}(\text{do}(\text{alice}, \text{inform}(\text{bob}, Q)), \text{s0}, \text{s1}) \wedge \\
\quad \forall_S\ \text{holds}(S, Q) \Leftrightarrow \\
\quad\quad \forall_{S2,S3,Q1} \\
\quad\quad\quad [S2 < S3 \leqslant S \wedge \\
\quad\quad\quad \text{occurs}(\text{do}(\text{bob}, \text{inform}(\text{alice}, Q1)), S2, S3)] \Rightarrow \\
\quad\quad\quad \forall_{S2A}\ \text{k_acc}(\text{alice}, S2, S2A) \Rightarrow \text{holds}(S2A, Q1)
\end{array}$$

4.2. Sample inferences

We next illustrate the inferential power of the above theory with three toy problems.

4.2.1. Sample inference 1

Given:

X.1: Sam knows in s0 that it will be sunny on July 4.

$$\text{holds}(\text{s0}, \text{know}(\text{sam}, \lambda(S)\ \forall_{SJ}\ S < SJ \wedge \text{time}(SJ) = \text{july4} \Rightarrow \text{holds}(SJ, \text{sunny})))$$

Expanding gives

$$\begin{aligned}&\forall_{S0A,SJA}\ [\text{k\_acc}(\text{sam}, \text{s0}, S0A) \wedge S0A < SJA \wedge \text{time}(SJA) = \text{july4}]\\ &\quad \Rightarrow \text{holds}(SJA, \text{sunny})\end{aligned}$$

X.2: In any situation, if it is sunny, then Bob can play tennis.

$$\forall_S\ \text{holds}(S, \text{sunny}) \Rightarrow \text{feasible}(\text{occurs}(\text{do}(\text{bob}, \text{tennis}), S))$$

X.3: Sam can always communicate with Bob.

$$\forall_{S1}\ \text{feasible}(\text{do}(\text{sam}, \text{communicate}(\text{bob})), S1).$$

Infer:

X.P: Sam knows that there is an action he can do (e.g., tell Bob that it will be sunny) that will cause Bob to know that he will be able to play tennis on July 4.

$$\begin{aligned}&\text{holds}(\text{s0}, \text{know}(\text{sam}, \lambda(S)\\ &\quad \exists_Z\ \text{feasible}(\text{do}(\text{sam}, Z), S) \wedge\\ &\qquad \forall_{S2A}\ \text{occurs}(\text{do}(\text{sam}, Z), S, S2A) \Rightarrow\\ &\qquad\quad \text{holds}(S2A, \text{know}(\text{bob}, \lambda(S2B)\\ &\qquad\qquad\qquad \forall_{S3B}\ S2B < S3B \wedge \text{time}(S3B) = \text{july4} \Rightarrow\\ &\qquad\qquad\qquad \text{feasible}(\text{do}(\text{bob}, \text{tennis}), S3B)))))\end{aligned}$$

Expanding gives

$$\begin{aligned}&\text{k\_acc}(\text{sam}, \text{s0}, S0A) \Rightarrow\\ &\exists_Z\ \text{feasible}(S0A, \text{do}(\text{sam}, Z)) \wedge\\ &\quad \forall_{S2A,S2B,S3B}\ [\text{occurs}(\text{do}(\text{sam}, Z), S0A, S2A) \wedge \text{k\_acc}(\text{bob}, S2A, S2B) \wedge\\ &\qquad\qquad S2B < S3B \wedge \text{time}(S3B) = \text{july4}] \Rightarrow\\ &\qquad\quad \text{feasible}(\text{do}(\text{bob}, \text{tennis}), S3B)\end{aligned}$$

Proof. By the comprehension axiom I.5 there is a fluent q1 that holds in any situation S just if it will be sunny on July 4 following S.

P.1: $\forall_S\ \text{holds}(S, \text{q1}) \Leftrightarrow [\forall_{S1}[S < S1 \wedge \text{time}(S1) = \text{july4}] \Rightarrow \text{holds}(S1, \text{sunny})]$.

Let z1 = inform(bob, q1). By axioms I.2, X.1, and X.3, do(sam, z1) is feasible in s0.

P.2: $\text{feasible}(\text{do}(\text{sam}, \text{z1}), \text{s0})$.

By axiom K.5, Sam knows in s0 that do(sam, z1) is feasible.

P.3: $\forall_{S0A}$ k_acc(sam, s0, $S0A) \Rightarrow$ feasible(do(sam, z1), $S0A$).

Let s0a be any situation such that k_acc(sam, s0, s0a). By P.3, there exists a situation s1a such occurs(do(sam, z1), s0a, s1a). Let s2a be any situation such that occurs(do(sam, z1), s0a, s2a). Let s2b be any situation such that k_acc(bob, s2a, s2b).

By Lemma 3.2, there exists s1b such that occurs(do(sam, z1), s1b, s2b) and holds(s1b, q1). Let s3b be any situation such that s2b $<$ s3b and time(s3b) $=$ july4. By T.8 and T.4, s1b $<$ s3b. By P.2, holds(s3b, sunny). By X.2, feasible(do(bob, tennis), s3b). Applying universal abstraction over s0a, s2a, s2b, and s3b and existential abstraction over z1 and s1a gives us formula X.P. □

4.3. Sample inference 2

Given:

Y.1: Bob confesses to Alice that he has cheated on her.

$$\begin{array}{l}\text{occurs(do(bob, inform(alice,}\\ \qquad \lambda(S)\ \exists_{S2,S3}\ S3 < S \wedge \text{occurs(do(bob, cheat)}, S2, S3))),\\ \qquad \text{s0, s1)}\end{array}$$

This expands to

$$\begin{array}{l}\exists_Q\ \text{occurs(do(bob, inform(alice}, Q)), \text{s0, s1)} \wedge\\ \forall_S\ \text{holds}(S, Q) \Leftrightarrow \exists_{S2,S3}\ S3 < S \wedge \text{occurs(do(bob, cheat)}, S2, S3)\end{array}$$

Y.2: Alice responds that Bob has never told her anything she did not already know.

As in sample representation 5, above, in expanded form this is:

$$\begin{array}{l}\exists_Q \text{occurs(do(alice, inform(bob}, Q)), \text{s1, s2)} \wedge\\ \quad \forall_S\ \text{holds}(S, Q) \Leftrightarrow\\ \quad\quad \forall_{S3,S4,Q1}\\ \quad\quad\quad [S3 < S4 \leqslant S \wedge \text{occurs(do(bob, inform(alice}, Q1)), S3, S4)] \Rightarrow\\ \quad\quad\quad \forall_{S3A}\ \text{k\_acc(alice}, S3, S3A) \Rightarrow \text{holds}(S3A, Q1)\end{array}$$

Y.P: Bob now knows that Alice had already known, before he spoke, that he had cheated on her.

$$\begin{array}{l}\text{holds(s2, know(bob,}\\ \quad \lambda(S2A)\ \exists_{S0A,S1A,Q1}\ S1A < S2A \wedge\\ \quad\quad \text{occurs(do(bob, inform(alice}, Q1)), S0A, S1A) \wedge\\ \quad\quad \text{holds}(S0A, \text{know(alice,}\\ \quad\quad\quad \lambda(S0B)\ \exists_{S3B,S4B}\ S4B < S0B \wedge\\ \quad\quad\quad\quad \text{occurs(do(bob, cheat)}, S3B, S4B))))).\end{array}$$

Expanding and rearranging gives:

$$\begin{aligned}
&\forall_{S2A}\ \text{k_acc}(\text{bob}, \text{s2}, S2A) \Rightarrow \\
&\quad \exists_{S0A,S1A,Q1}\ S1A < S2A \wedge \text{occurs}(\text{do}(\text{bob}, \text{inform}(\text{alice}, Q1)), S0A, S1A) \wedge \\
&\qquad [\forall_{S0B}\ \text{k_acc}(\text{alice}, S0A, S0B) \Rightarrow \\
&\qquad \exists_{S3B,S4B}\ S4B < S0B \wedge \text{occurs}(\text{do}(\text{bob}, \text{cheat}), S3B, S4B)]
\end{aligned}$$

Proof. Let q1 be the content of Bob's statement in Y.1, and let q2 be the content of Alice's statement in Y.2. By axiom I.5, both these fluents exist.

By K.4, Bob knows in s2 that he has informed Alice of q1.

Q.1: $\forall_{S2A}$ k_acc(bob, s2, $S2A$) $\Rightarrow$
$\exists_{S0A,S1A}\ S1A < S2A \wedge \text{occurs}(\text{do}(\text{bob}, \text{inform}(\text{alice}, \text{q1})), S0A, S1A)$.

By Lemma 3.2, Bob knows in s2 that q2 held when Alice started to speak.

Q.2: k_acc(bob, s2, $S2A$) $\Rightarrow$
$\exists_{S1A}\ \text{occurs}(\text{do}(\text{alice}, \text{inform}(\text{bob}, \text{q2})), S1A, S2A) \wedge \text{holds}(S1A, \text{q2})$.

Let s2a be any situation such that k_acc(bob, s2, s2a), and let s1a be a corresponding value of $S1A$ satisfying Q.2. Then holds(s1a, q2).

By definition of q2, we have that in s1a, whenever Bob had previously told Alice anything ($Q3$), she had already known it.

Q.3: $\forall_{S3,S4,Q3}\ [S3 < S4 \leqslant \text{s1a} \wedge \text{occurs}(\text{do}(\text{bob}, \text{inform}(\text{alice}, Q3)), S3, S4)] \Rightarrow$
$\forall_{S3A}\ \text{k\_acc}(\text{alice}, S3, S3A) \Rightarrow \text{holds}(S3A, Q1)$.

By K.4 and Y.3, Bob knows in s1 that he has informed Alice of q1.

Q.4: $\forall_{S1A}$ k_acc(bob, s1, $S1A$) $\Rightarrow$
$\exists_{S0A}\ \text{occurs}(\text{do}(\text{bob}, \text{inform}(\text{alice}, \text{q1})), S0A, S1A)$.

In particular, therefore, Q.4 is true of $S1A = \text{s1a}$.

Q.5: $\exists_{S0A}\ \text{occurs}(\text{do}(\text{bob}, \text{inform}(\text{alice}, \text{q1})), S0A, \text{s1a})$.

Let s0a be a situation satisfying Q.5. Applying Q.3, with $S3 \to \text{s0z}$, $S4 \to \text{s1a}$, and $Q3 \to \text{q1}$, gives

Q.6. $\forall_{S0B}\ \text{k\_acc}(\text{alice}, \text{s0a}, S0B) \Rightarrow \text{holds}(S0B, \text{q1})$. $\square$

Applying the definition of q1, we get the desired result.

4.4. Sample inference 3

Given:

Z.1: Anne does not know that she has a sister.

$$\neg\text{holds}(\text{s0}, \text{know}(\text{anne}, \lambda(S)\ \exists_Y\ \text{holds}(S, \text{sister}(Y, \text{anne}))))$$

This expands to

$$\neg[\forall_{S0A}\ \text{k\_acc}(\text{anne}, \text{s0}, S0A) \Rightarrow \exists_Y\ \text{holds}(S0A, \text{sister}(Y, \text{anne}))]$$

Z.2: Anne knows that, if she had a sister, someone would have told her about him.

$$\begin{array}{l}\text{holds}(\text{s0}, \text{know}(\text{anne},\\ \quad \lambda(S)\ \forall_Y\ \text{holds}(S, \text{sister}(Y, \text{anne})) \Rightarrow\\ \quad\quad \exists_{S1,S2,AS}\ S2 < S \wedge \text{occurs}(\text{do}(AS, \text{inform}(\text{anne}, \text{sister}(Y, \text{anne}))), S1, S2)))\end{array}$$

Expanding and rearranging,

$$\begin{array}{l}\forall_{S0A}\ \text{k\_acc}(\text{anne}, \text{s0}, S0A) \Rightarrow\\ \quad \forall_Y\ \text{holds}(S0A, \text{sister}(Y, \text{anne})) \Rightarrow\\ \quad\quad \exists_{S1A,S2A,AS}\ S2A \leqslant S0A \wedge\\ \quad\quad\quad \text{occurs}(\text{do}(AS, \text{inform}(\text{anne}, \text{sister}(Y, \text{anne}))), S1A, S2A)\end{array}$$

Z.3: Sisterhood is forever.

$$S0 < S1 \wedge \text{holds}(S0, \text{sister}(X, Y)) \Rightarrow \text{holds}(S1, \text{sister}(X, Y))$$

Infer: Anne knows that she has no sister.

$$\text{holds}(\text{s0}, \text{know}(\text{anne}, \lambda(S)\ \neg\exists_Y\ \text{holds}(S, \text{sister}(Y, \text{anne}))))$$

Expands to:

Z.4: $\forall_{S0A}\ \text{k\_acc}(\text{anne}, \text{s0}, S0A) \Rightarrow \neg\exists_Y\ \text{holds}(S0A, \text{sister}(Y, \text{anne}))$.

Note: This is a monotonic variant of the "auto-epistemic" inference [30].

Proof by contradiction. Suppose that Z.4 is false and Anne does not know that she does not has a sister—in other words, as far as she knows she might have a sister.

R.1: $\exists_{S0A,Y}\ \text{k\_acc}(\text{anne}, \text{s0}, S0A) \wedge \text{holds}(S0A, \text{sister}(Y, \text{anne}))$.

Let sb and yb be values satisfying R.1. Thus k_acc(anne, s0, sb) and holds(sb, sister(yb, anne)). By Z.2, in sb someone would have already told her that she had a sister.

R.2: $\exists_{S1A,S2A,AS}\ S2A \leqslant \text{sb} \wedge \text{occurs}(\text{do}(AS, \text{inform}(\text{anne}, \text{sister}(\text{yb}, \text{anne}))), S1A, S2A)$.

By Lemma 3.2, Anne would know in sb that she had previously had a sister.

R.3: $\forall_{SC}\ \text{k\_acc}(\text{anne}, \text{sb}, SC) \Rightarrow$
$\exists_{S1C}\ S1C < SC \wedge \text{holds}(S1C, \text{sister}(\text{yb}, \text{anne}))$.

Let s0x be any situation such that k_acc(anne, s0, s0x). By K.2 and K.3 k_acc(anne, sb, s0x). By R.3 and X.3, holds(s0x, sister(yb, anne)). Applying universal abstraction to s0x we have

R.4: $\forall_{S0X}\ \text{k\_acc}(\text{anne}, S0, S0X) \Rightarrow \text{holds}(S0X, \text{sister}(\text{yb}, \text{anne}))$.

But this contradicts X.1. □

5. Paradox

The following Russell-like paradox seems to threaten our theory:[10]

Paradox: Let Q be a fluent. Suppose that over interval $[S0, S1]$, agent a1 carries out the action of informing a2 that Q holds. Necessarily, Q must hold in $S0$, since agents are not allowed to lie (axiom I.2). Let us say that this communication is *immediately obsolete* if Q no longer holds in $S1$. For example, if it is raining in s0, the event of a1 telling a2 that it is raining occurs over [s0, s1], and it has stopped raining in s1, then this communication is immediately obsolete. Now let us say that a1 has "misled" a2 in S if S is the end of an immediately obsolete communication. (There is no suggestion intended here, of course, that a2 has any false beliefs.) Since "a1 having misled a2" is a property of a situation, by the comprehension axiom it should be definable as a fluent. Symbolically,

$$\text{holds}(S, \text{misled}(A1, A2)) \equiv$$
$$\exists_{Q,A1,A2,S0}\ \text{occurs}(\text{do}(A1, \text{inform}(A2, Q)), S0, S) \wedge \neg\text{holds}(S, Q)$$

Now, suppose that, as above, in s0 it is raining; from s0 to s1, a1 tells a2 that it is raining; and in s1 it is no longer raining and a1 knows that it is no longer raining. Then a1 knows that "misled(a1, a2)" holds in s1. Therefore, (axiom I.2) it is feasible for a1 to tell a2 that "misled(a1, a2)" holds in s1. Suppose that, from s1 to s2, a1 informs a2 that "misled(a1, a2)" holds. The question is now, does "misled(a1, a2)" hold in s2? Well, if it does, then what was communicated over [s1, s2] still holds in s2, so "misled(a1, a2)" does not hold; but if it does not, then what was communicated no longer holds, so "misled(a1, a2)" does hold in s2.

The flaw in this argument is that it presupposes the independence axiom WRONG.1 (p. 93) that we rejected earlier. The argument presumes that if fluent $Q1 \neq Q2$, and do($A1$, inform($A2$, $Q1$)) occurs from s1 to s2, then do($A1$, inform($A2$, $Q2$)) does not occur. (Our English description of the argument used the phrase "what was communicated between s1 and s2“, which presupposes that there was a unique content that was communicated.) But axiom I.4 asserts that many different fluents are communicated in the same act. Therefore, the argument collapses.

In particular, as we shall show, the clock time (in the sense of "the number of situations that have elapsed since the start of time") is always common knowledge among all agents (Theorem 3, Appendix A). Now, let q1 be any fluent, and suppose that occurs(do(a1, inform(a2, q1)), s1, s2). Let t1 = time(q1) and let q2 be the fluent defined by the formula

$$\forall_S\ \text{holds}(S, \text{q2}) \Leftrightarrow \text{holds}(S, \text{q1}) \wedge \text{time}(S) = \text{t1}$$

By assumption, it is shared knowledge between a1 and a2 that holds(s1, q2) $\Leftrightarrow$ holds(s1, q1). Hence, by axiom I.4, occurs(do(a1, inform(a2, q2)), s1, s2). But by construction q2 does not hold in s1; hence the occurrence of do(a1, inform(as, q2)) from s1 to s2 is immediately obsolete. Therefore "misled(a1, a2)" holds *any* time a1 communicates with a2.

[10] The comprehension axiom in itself, without the "inform" acts, does not lead to Russell's paradox, because a *fluent* is being defined as in terms of a property of *situations*, so that there is no circularity. Formally, we will construct a set of situations, and then use the standard Zermelo–Fraenkel separation axiom to define a fluent as a subset. See Lemma A.21, p. 136.

Changing the definition of misled to use the universal quantifier, thus:

$$\text{holds}(S, \text{misled}(A1, A2)) \equiv$$
$$\forall_{Q,A1,A2}\ \text{occurs}\ (\text{do}(A1, \text{inform}(A, Q)), S0, S) \wedge \neg\text{holds}(S, Q)$$

does not rescue the contradiction. One need only change the definition of q2 above to be

$$\forall_S\ \text{holds}(S, \text{q2}) \Leftrightarrow \text{holds}(S, \text{q1}) \vee \text{time}(S) \neq \text{t1}$$

Clearly, the new definition of “misled(a1, a2)” *never* holds after any informative act.

Of course, if we extend the theory to include the underlying locutionary act, then this paradox may well return, as the locutionary act that occurs presumably is unique. However, as the content of a locutionary act is a quoted string, we can expect to have our hands full of paradoxes in that theory; this “misled” paradox will not be our biggest problem [31].

6. Unexpected hanging

The well-known paradox of the unexpected hanging (also known as the surprise examination) [16,34] can be formally expressed in our theory; however, the paradox does not render the theory inconsistent. (The analysis below is certainly *not* a philosophically adequate solution to the paradox, merely an explanation of how our particular theory manages to side-step it.)

The paradox can be stated as follows:

> A judge announces to a prisoner, “You will be hung at noon within 30 days; however, that morning you will not know that you will be hung that day”. The prisoner reasons to himself, “If they leave me alive until the 30th day, then I will know that morning that they will hang me that day. Therefore, they will have to kill me no later than the 29th day. So if I find myself alive on the morning of the 29th day, I can be sure that I will be hung that day. So they will have to kill me no later than the 28th day .... So they cannot kill me at all!”
>
> On the 17th day, they hung him at noon. He did not know that morning that he would be hung that day.

Let kill_today be the fluent that the prisoner will be hung today. Then the judge’s statement can be represented as follows:

$$\begin{aligned}
&\text{occurs}(\text{do}(\text{judge}, \text{inform}(\text{prisoner},\\
&\quad \lambda(S)\ \forall_{SX}\ [S < SX \wedge \text{time}(SX) = \text{time}(S) + 31] \Rightarrow\\
&\qquad \exists_{SH}\ S < SH < SX \wedge \text{holds}(SH, \text{kill_today}) \wedge\\
&\qquad\quad \neg\text{holds}(SH, \text{know}(\text{prisoner}, \text{kill_ today})))),\\
&\text{s0}, \text{s1})
\end{aligned}$$

Expanding and rearranging, this becomes

$$\begin{array}{l} \exists_Q \text{ occurs}(\text{do}(\text{judge}, \text{inform}(\text{prisoner}, Q)), \text{s0}, \text{s1}) \wedge \\ \quad \forall_S \text{ holds}(S, Q) \Leftrightarrow \\ \qquad \exists_{SH,SHA}\ S < SH < SX \wedge \text{holds}(SH, \text{kill\_today}) \wedge \\ \qquad\quad \text{k\_acc}(\text{prisoner}, SH, SHA) \wedge \neg\text{holds}(SHA, \text{kill\_today}) \end{array}$$

We further posit the axioms that the prisoner has never been killed before s0, and that if an agent has never been killed, he knows that he has never been killed.[11]

$$\neg\exists_S\ S < \text{s0} \wedge \text{holds}(S, \text{kill\_today}).$$

$$\begin{array}{l} \forall_{S2}\ [\forall_{S1}\ S1 < S2 \Rightarrow \neg\text{holds}(S1, \text{kill\_today})] \Rightarrow \\ \text{holds}(S2, \text{know}(\text{prisoner}, \lambda(S)\ \forall_{S1}\ S1 < S \Rightarrow \neg\text{holds}(S1, \text{kill\_today}))) \end{array}$$

Let UH^{lang} be the judge's statement in English and let UH^{logic} be the fluent Q that the judge communicates. Let "kill(K)" be the proposition that the prisoner will be killed no later than the Kth day. It would appear that UH^{lang} is true; that the judge knows that in s0 that it is true, and that UH^{logic} means the same as UH^{lang}. By Axiom I.2, if the judge knows that UH^{logic} holds in s0, then he can inform the prisoner of it. How, then, does our theory avoid contradiction?

The first thing to note is that the prisoner *cannot* know UH^{logic}. There is simply no possible worlds structure in which the prisoner knows UH^{logic}. The proof is exactly isomorphic to the sequence of reasoning that prisoner goes through. Therefore, by Lemma 3.2 above, the judge cannot inform the prisoner of UH^{logic}; if he did, the prisoner would know it to be true.

The critical point is that there is a subtle difference between UH^{lang} and UH^{logic}. The statement UH^{lang} asserts that the prisoner *will* not know kill_today—this means even *after* the judge finishes speaking. In our theory, however, one can only communicate properties of the situation at the beginning of the speech act and there is no way to refer to what *will* happens as distinguished from one *could* happen. So what UH^{logic} asserts is that the prisoner will not know kill_today *whatever* the judge decides to say or do in s0.

In fact, it is easily shown that either [the judge does not know in s0 that UH^{logic} is true], or [UH^{logic} is false]. It depends on what the judge knows in s0. Let us suppose that in s0, it is inevitable that the prisoner will be killed on day 17 (the executioner has gotten irrevocable orders). There are two main cases to consider.

- *Case* 1: All the judge knows is kill(K), for some $K > 17$. Then the most that the judge can tell the prisoner is kill(K). In this case, UH^{logic} is in fact true in s0, but the judge does not know that it is true, because as far as the judge knows, it is possible that (a) he will tell the prisoner kill(K) and (b) the prisoner will be left alive until the Kth day, in which case the prisoner would know kill_today on the morning of the Kth day.

[11] In S5, it is a logical consequence of this axiom that if he has been killed, he knows he has been killed; but that is beyond the scope of this paper.

- *Case* 2: The judge knows kill(17). In that case, UH^{logic} is not even true in s0, because the judge has the option of telling the prisoner kill(17), in which case the prisoner will know kill_today on the morning of the 17th day.

Again, we do not claim that this is an adequate solution to the philosophical problem, merely an explanation of how our formal theory manages to remain consistent and sidestep the paradox. In fact, in the broader context the solution is not at all satisfying, for reasons that may well become serious when the theory is extended to be more powerful. There are two objections. First, the solution depends critically on the restriction that agents cannot talk about what *will* happen as opposed to what *can* happen; in talking about the future, they cannot take into account their own decisions or commitments about what they themselves are planning to do. One can extend the outer theory so as to be able to *represent* what will happen—in [10], we essentially do this—but then the comprehension axiom I.5 must be restricted so as to exclude this from the scope of fluents that can be the content of an "inform" act. We do not see how this limitation can be overcome.

The second objection is that it depends on the possibility of the judge telling the prisoner kill(17) if he knows this. Suppose that we eliminate this possibility? Consider the following scenario: The judge knows kill(17), but he is unable to speak directly to the prisoner. Rather, he has the option of playing one of two tape recordings; one says "kill(30)" and the other says UH^{logic}. Now the theory is indeed inconsistent. Since the prisoner cannot know UH^{logic} it follows that the judge cannot inform him of UH^{logic}; therefore the only thing that the judge can say is "kill(30)". But in that case, the formula UH^{logic} is indeed true, and the judge knows it, so he should be able to push that button.

To axiomatize this situation we must change axiom I.2 to assert that the only possible inform acts are kill(30) and UH^{logic}.

Within the context of our theory, it seems to me that the correct answer is "So what?" Yes, you can set up a Rube Goldberg mechanism that creates this contradiction, but the problem is not with the theory, it is with the axiom that states that only these two inform acts are physically possible.

(Those readers, if any, who work through the proof of Theorem 1 in Appendix A may wonder what prevents this constraint from being incorporated into the construction of u-situations. After all, all that this amounts to is drastically restricting the class of "inform" acts that are added on. The answer is that which of the "inform" acts are allowed to exist now depends on the interpretation of a formula in the extended language, and that therefore the construction now involves a vicious cycle. See further the comments on Lemma A.21.)

In a wider context, though, this answer will not serve. After all, it is physically possible to create this situation, and in a sufficiently rich theory of communication, it will be provable that you can create this situation. However, such a theory describing the physical reality of communication must include a theory of locutionary acts; i.e., sending signals of quoted strings. As mentioned above such a theory will run into *many* paradoxes; this one is probably not the most troublesome.

7. Consistency

Two paradoxes have come up, but the theory has side-stepped them both. How do we know that the next paradox won't uncover an actual inconsistency in the theory? We can eliminate all worry about paradoxes once and for all by proving that the theory is consistent. We do this by constructing a model satisfying the theory. More precisely, we construct a fairly broad class of models, establishing (informally) that the theory is not only consistent but does not necessitate any weird or highly restrictive consequences. (Just showing soundness with respect to a model or even completeness is not sufficient for this. For instance, if the theory were consistent only with a model in which every agent was always omniscient, and inform acts were therefore no-ops, then the theory would be *consistent* but not of any interest.)

As usual, establishing soundness has three steps: defining a model, defining an interpretation of the symbols in the model, and establishing that the axioms are true under the interpretation.

Our class of models is (apparently) more restrictive than the theory;[12] that is, the theory is not complete with respect to this class of models. The major additional restrictions in our model are:

I. Time must be discrete. We believe that this restriction can be lifted with minor modifications to the axioms, but this is beyond the scope of this paper. We hope to address it in future work.
II. Time must have a starting point; it cannot extend infinitely far back. It would seem to be very difficult to modify our proof to remove this constraint; at the current time, it seems to depend on the existence of highly non-standard models of set theory.
III. A knowledge accessibility link always connects two situations whose time is equal, where "time" measure the number of clock ticks since the start. In other words, all agents always have common knowledge of the time. In a discrete structure, this is a consequence of the axiom of memory. Therefore, it is not, strictly speaking, an additional restriction; rather, it is a non-obvious consequence of restriction (I). If we extend the construction to a non-discrete time line, some version of this restriction must be stated separately.

There are also more minor restrictions; for example, we will define shared knowledge to be the true transitive closure of knowledge, which is not expressible in a first-order language.

Theorem 1 below states that the axioms in this theory are consistent with essentially any physical theory that has a model over discrete time with a starting point state and physical actions.

Definition 1. A *physical language* is a first-order language containing the sorts "situations", "agents", "physical actionals", "physical actions", "physical fluents", and "clock

[12] The only way to be sure that the theory is more general than the class of models is to prove that it is consistent with a broader class of models.

times"; containing the non-logical symbols, "$<$", "do", "occurs", "holds", "time", and "communicate"; and excluding the symbols, "k_acc", "inform", and "sk_acc".

Definition 2 (*This is Definition* A.6 *of Appendix* A). Let $\mathcal{L}$ be a physical language, let $\mathcal{T}$ be a theory over $\mathcal{L}$. $\mathcal{T}$ is an *acceptable physical theory* (i.e. acceptable for use in Theorem 1 below) if there exists a model $\mathcal{M}$ and an interpretation $\mathcal{I}$ of $\mathcal{L}$ over $\mathcal{M}$ such that the following conditions are satisfied:

(1) $\mathcal{I}$ maps the sort of clock times to the positive integers, and the relation $T1 < T2$ on clock times to the usual ordering on integers.
(2) Axioms T.1–T.9 in Table 1 are true in $\mathcal{M}$ under $\mathcal{I}$.
(3) Theory $\mathcal{T}$ is true in $\mathcal{M}$ under $\mathcal{I}$.
(4) The theory is consistent with the following constraint: In any situation S, if any communication act is feasible, then arbitrarily many physically indistinguishable communication acts are feasible.
(5) If α is a predicate symbol in $\mathcal{L}$ with more than one situational argument, then $\alpha(X_1 \dots X_k)$ holds only if all the situations among $X_1 \dots X_k$ are ordered with respect to $<$. (Note that this condition holds both when α is "$<$" and α is "occurs".) If $\beta(X_1 \dots X_k)$ is a function symbol, then the above condition holds for the relation $X_{k+1} = \beta(X_1 \dots X_k)$.

Condition (4) no doubt seems complex, strange, and restrictive. But in fact any physical model can be easily transformed into one satisfying this condition: take the original model and, wherever a communicative act occurs, make an infinite number of identical copies of the subtree following the branch where the act occurs. Moreover, most reasonable physical theories $\mathcal{T}$ will accept this transformation, or can be straightforwardly transformed into theories that will accept this transformation. In fact, therefore, condition (4) is not a substantial restriction on $\mathcal{T}$. The reason it is needed is that, without this condition, the physical theory could include an axiom like, "In any situation S there is only one situation $S1$ such that occurs(do(*AS*, communicate(*AH*), S, $S1$)" whereas our theory demands that there must exist many such situations corresponding to the different informative acts possible in S.

Condition (5) is a technical one needed for the proof. We do not know of any reasonable causal theories that contain predicates that do not satisfy this condition and that cannot be defined in terms of simpler predicates that satisfy this condition. We do not know whether (5) is necessary for the conclusion of Theorem 1 to hold; however, we have not been able to construct a proof without it.

(The KR-2004 paper claims that condition (4) can be stated in a first-order axiom schema. This is in error. More precisely, I have not found any first-order axiom schema that can be used to instantiate condition (4) that I can prove to be sufficient for the theorem below.)

Theorem 1. *Let $\mathcal{T}$ be an acceptable physical theory, and let $\mathcal{U}$ be $\mathcal{T}$ together with axioms* K.1–K.8 *and* I.1–I.5. *Then $\mathcal{U}$ is consistent.*

The proof of Theorem 1 is given in Appendix A.

It is possible to strengthen Theorem 1 by adding in domain-specific axioms of knowledge acquisition and the associated frame axiom over accessibility relation, as described in Section 3.5, plus conditions on the initial knowledge and ignorance of the agents. Specifically, we have the following theorem:

Theorem 2. *Let $\mathcal{T}$ be an acceptable physical theory, and let $\mathcal{U}$ be the union of*:

A. $\mathcal{T}$.
B. *Axioms* K.1–K.7 *and* I.1–I.5.
C. *A collection of domain-specific knowledge acquisition axioms of the form specified in Section* 3.5.
D. *The frame axiom* I.6 *associated with the axioms in* (C).
E. *Any set of axioms $\mathcal{K}$ specifying knowledge or ignorance at time* 0 *as long as*:
   i. *The axioms in $\mathcal{K}$ do not refer to any situations of time later than* 0.
   ii. *The axioms in $\mathcal{K}$ are consistent with $\mathcal{T}$, axioms* K.1–K.3, K.5 (*as regards knowing the feasibility of actions at time* 0); *and the axioms in* (C).

Then $\mathcal{U}$ is consistent.

In Appendix A, we sketch how the proof of Theorem 1 is modified to give a proof of Theorem 2.

8. Related work

The theory presented here was originally developed as part of a larger theory of multi-agent planning [10]. That theory includes requests as speech acts as well as informative speech acts. However, our analysis of informative acts there was not as deep or as extensive in scope.

As far as we know, this is the first attempt to characterize the content of communication as a first-order property of possible worlds. Morgenstern [31] develops a theory in which the content of communication is a string of characters. A number of BDI models incorporate various types of communication. The general BDI model was first proposed by Cohen and Perrault [4]; within that model, they formalized illocutionary acts such as “Request” and “Inform” and perlocutionary acts such as “Convince” using a STRIPS-like representation of preconditions and effects on the mental states of the speaker and hearer. Cohen and Levesque [5] extend and generalize this work using an full modal logic of time and propositional attitudes. Here, speech acts are *defined* in terms of their effects; a request, for example, is any sequence of actions that achieves the specified effect in the mental state of the hearer.

Update logic (e.g. [2,33]) combines dynamic logic with epistemic logic, introducing the dynamic operator $[A!]\phi$, meaning “ϕ holds after A has been truthfully announced”. The properties of this logic have been extensively studied. Baltag et al. [1] extend this logic to allow communication to a subset of agents, and to allow “suspicious” agents. Colombetti [6] proposes a *timeless* modal language of communication, to deal with the interaction of intention and knowledge in communication. Parikh and Ramanujam [32] present a theory of *messages* in which the meaning of a message is interpreted relative to a protocol.

There is a large literature on the applications of modal logics of knowledge to a multi-agent systems. For example, Sadek et al. [37] present a first-order theory with two modal operators $\mathrm{B}_i(\phi)$ and $\mathrm{I}_i(\phi)$ meaning "Agent i believes that ϕ" and "Agent i intends that ϕ" respectively. An inference engine has been developed for this theory, and there is an application to automated telephone dialogue that uses the inference engine to choose appropriate responses to requests for information. However, the temporal language associated with this theory is both limited and awkward; it seems unlikely that the theory could be applied to problems involving multi-step planning. (The dialogue application requires only an immediate response to a query.)

The multi-agent communication languages KQML [13] and FIPA [14] provide rich sets of communication "performatives". KQML was never tightly defined [40]. FIPA has a formal semantics defined in terms of the theory of Sadek et al. [37] discussed above. However, the content of messages is unconstrained; thus, the semantics of the representation is not inherently connected with the semantics of the content, as in our theory.

Other modal theories of communication, mostly propositional rather than first-order, are discussed in [23,35,41].

9. Fagin, Halpern, Moses, and Vardi

The theory of runs and messages, developed by Fagin, Halpern, Moses, and Vardi (FHMV) in their book *Reasoning about Knowledge* [12] and many papers, presents a constructive model of a system of agents. Each agent is characterized as a infinite sequence. The state of agent A at time T is the prefix of the first T elements of the corresponding sequence. The global state of the system at time T is the tuple of the states of all the agents at time T. Two global system states Q1 and Q2 are knowledge accessible relative to A if the state of A is the same in Q1 and Q2. A message is an event that modifies the state of the sender when it is sent and the state of the recipient when received. There is a protocol that governs under what circumstances a sender may send a specified message. Messages may be given a semantics, and agents can be prohibited from sending messages that they know to be false.

Thus, the FHMV theory deals with much the same issues as our theory, and arrives at many of the same rules: axioms K.1–K.3, K.7, and K.8 are valid in all FHMV models, and FHMV have extensively studied classes of models in which axioms K.4, K.6, I.3, I.6, and the forward implication in I.2 are valid.

Nonetheless there are many major differences between FHMV and our theory. We divide these differences for the most part into three categories: differences in purpose, differences in the model, and differences in the representation language. These three categories interact strongly.

9.1. Differences in purpose

The central objective of FHMV is to establish a theory for characterizing distributed systems in terms of the "knowledge" of the components and the communications between them. Such a theory can be used as the foundation for the formal analysis of such systems; e.g. proving that a given class of systems is safe, in some sense; that a specified protocol achieves a specified goal; that a given state of knowledge is inevitable or unattainable; and so on. These proofs might be carried out automatically by reasoning in terms of the formal language, but more often FHMV seem to be thinking about proofs carried out by human reasoners reasoning directly about the model. In most cases, the construction of the model is the critical issue; the definition of a formal language and statement of axioms is secondary, or peripheral. Indeed, in [20] Halpern and Vardi argue strongly in favor of a model-based as opposed to axiom-based approach to automated reasoning.

By contrast, our central objective here is, primarily, to define a formal language capable of representing a wide range of statements about knowledge in commonsense domains, and, secondarily, to demonstrate the power of this language by formulating axioms sufficient to justify commonsense inferences. Ultimately, the language and axioms could serve as the basis for a representation and rule set in a symbolic knowledge base. The model is secondary, as is evidenced by the fact that it is only described in the appendix; it is constructed only in order to enable us to prove that our representation is coherent and our theory is consistent. (For that reason, the inelegance, not to say ugliness, of our model does not much matter.)

This difference also underlies our different attitudes toward completeness proofs. FHMV construct models that are elegant and interesting in themselves; it is therefore a worthwhile enterprise looking for axiom sets that characterize them exactly. But our model is constructed out of scotch tape and toothpicks to fit the axioms: what would be gained by finding a complete axiom set, even if it were possible? After all, since the theory is first-order and consistent, we can be sure that there exists a class of models with respect to which the theory is complete; namely, the class of all models satisfying the theory.

Another difference is that FHMV are much more interested in the properties of communication itself and communication channels, and have studied in depth the properties of systems with unreliable channels or with unknown delays. By contrast, we have been content to deal only with the case of direct speech, or, more generally, communication across a reliable channel of fixed delay.

9.2. Differences in model

The key difference between the two models might, at first glance, seem to be a rather technical one: FHMV uses a linear model of time[13] whereas we use a branching model of time. But that difference has many ramifications. In a linear model of time, one cannot speak of an actor having options of many different possible actions. Therefore, it is not

[13] There are a few exceptions; [18,19,27] consider FHMV models with branching time. However, even in these, no axioms are given that require that there ever exists more than one branch in a situation; that is, these theories *permit* time to branch but do not *require* it.

possible in the FHMV model to reason about what an agent *can* accomplish or communicate. Inferences such as sample inference 3 (that Sam can cause Bob to know that he will be able to play tennis) and axioms such as the forward implication of I.2 (that *AS* can inform *AH* of anything that *AS* knows to be true) are not merely invalid in the FHMV model; it is essentially impossible to formulate them in that setting. Indeed, almost all the predictive theorems in the FHMV theory are universal, asserting that a system must attain particular conditions, or cannot attain them; there are few existential theorems, asserting that a system can attain a particular condition. The comprehension axiom over fluents, in this setting, can be made true, but is essentially irrelevant; since any particular system contains only a restricted set of messages, the only fluents that need to exist are those that are the content of these messages.

For that reason, the FHMV model cannot be applied to automated planning under the usual logical analysis. The usual logical analysis of the planning problem of states that a deterministic[14] plan P correctly achieves goal G starting in situation $S0$, if (1) P is feasible starting in $S0$; that is, there exists an $S1 > S0$ such that P is executed over $[S0, S1]$; and (2) for all such $S1$, G is achieved over $[S0, S1]$. But in a linear model of time, (1) can never be true of two alternative but mutually exclusive plans.

Another difference in the model, reflecting to FHMV's interest in communication channels, is that where we have a single action "do(*AS*, inform(*AH*, Q))" which involves both the speaker and the hearer, FHMV separate this into two parts: one agent sends a message, then later another agent receives it. The FHMV model is much more general.

9.3. Differences in formal language

There are many differences between the FHMV formal language and our formal languages. To some extent, this reflects the difference in the model; to some extent it reflects the difference in purpose; to some extent it is a matter of personal preference in representation style. The representational choices all interact, which makes it difficult to separate out the different motivations behind the different choices. Among the most conspicuous differences are:

- FHMV use modal languages of time and knowledge where we use a first-order language. The modal logic formulation has the advantage of supporting interesting theorems about computability in the case where the base language is propositional.
- FHMV very rarely use an explicit representation of actions and events. They occasionally raise the possibility of using a dynamic logic, in which there is a modal operator corresponding to each action.
- The content of a communication is not a first-order entity in FHMV. Indeed, there is no representation of a message whose form reflects the meaning of the message; the meanings of message are set by a meta-level operator σ.
- Agents enter the formal language of FHMV primarily as subscripts on the modal operator. It is therefore not possible to *quantify* over agents.

[14] The logical analysis of non-deterministic plans such as partially ordered plans is more complex, but also require non-linear models of time [3].

- In general, FHMV aim at a very spare formal language; since our objective is to maximize expressivity, we tend to aim at a very rich one.

9.4. Other differences

One final difference: FHMV like to study, not a one single theory at a time, but a sheaf of variant theories, whereas we have presented a single theory. This difference, I think, is mostly a stylistic difference in research method, and is not closely related to any of the other differences.

Indeed, FHMV present a set of taxonomic categorizations of different theories of knowledge and communication [11]. In terms of that taxonomy, our theory has the following characteristics:

- The system is synchronous.
- Knowledge is cumulative.
- The environment does not determine the initial state of the agents.
- The system is not required to be history independent.
- Process state transitions are not independent of the environment, or of the initial environment.
- The system is not deterministic.
- The primitive propositions are not determined either by the current global state or by the initial global state.
- Neither the primitive propositions nor the class of agents is required to be finite.

10. Conclusions and open problems

We have developed a theory of communications which allows the content of an informative act to include quantifiers and logical operators and to refer to physical states, events including other informative acts, and states of knowledge; all these in the past, present, or possible futures. We have proven that this theory is consistent, and compatible with a wide range of physical theories. We have examined how the theory avoids two potential paradoxes, and discussed how these paradoxes may pose a danger when these theories are extended. Elsewhere [10] we have shown that the theory can be integrated with a similarly expressive theory of multi-agent planning.

The major technical problem that follows naturally on this work is to find ways to relax the limitations enumerated in Section 1 while preserving the consistency of the theory. Let us discuss what is involved here a little.

Two related restrictions are particularly significant in terms of limiting the scope of applications of this theory: first, that the sender *AS* knows when a communication has been received (a consequence of axiom K.5) and that the hearer knows when the communication was sent (a consequence of axiom I.3). To relax this restriction, it would be necessary, as in FHMV and similar theories, to separate the action "do(*AS*, inform(*AH*, *Q*))" into an action of sending a message and an exogenous event of receiving it. The difficulty is that we have not found a reasonable reformulation of axiom I.4 in a way that we can prove avoids paradox.

The restriction that the sender and recipient know each other is one that, in practice, is often enough violated, and it would certainly be interesting to relax this. If you relax this condition, then a timed communication (i.e., one satisfying I.4) gives rise to *anonymous shared knowledge*. That is, the speaker and hearer know that they share the knowledge of the content; they just do not know who they are sharing the knowledge with. (Or one knows and the other does not.) This is analogous to common knowledge among non-rigid sets [12, Section 6.4] but the different setting here raises different issues.

The restriction to discrete time obviously impedes the integration of this theory with physical theories that use continuous time. The problem is that the construction of the model in our consistency proof is inherently iterative over time, and there does not seem to be any easy way to modify this iterative structure. The proof will work if one makes strong assumptions about the discreteness of communicative acts; e.g., one posits that it is only physically possible to begin a communication in a situation whose clock time is a non-negative integer. It is conceivable that such a theory would suffice for most applications; one would have to look over examples of reasoning that integrate continuous physical reasoning with communication, which I have not yet done. I would conjecture that axioms K.1–K.7 and I.1–I.6 are in fact consistent with a continuous model of time, without modification, and without the need to impose strong conditions on the physics of communication, but I am certainly far from a proof.

Other, more far-reaching, problems include:

- Our work on integrating the theory here with a theory of planning [10] involves some rather restrictive constraints on the protocols between agents. We would like to study how the theory can be modified to weaken these.
- To my mind, the brass ring in this field would be to integrate the above theory of illocutionary acts, which describes the content of communications, with a theory of locutionary acts, which would describe the form of communications. Achieving a theory that is both general and consistent would be a major accomplishment.

Acknowledgement

The work described here comes out of and builds upon a project done in collaboration with Leora Morgenstern, stemming from a benchmark problem that she proposed. Thanks also to the reviewers for suggestions and corrections.

Appendix A. Proof of Theorem 1

This appendix contains a proof of Theorem 1. Specifically, we prove that if $\mathcal{T}$ is a physical theory over integer-valued time satisfying a few, not very restrictive, constraints, then $\mathcal{T}$ is consistent with our axioms of knowledge and of communication.

Outline of appendix: in Section A.1 we give a formal definition of what we mean by a physical theory. In Section A.2, we show how a model of a physical theory can be extended to incorporate knowledge relations and informative actions. In Section A.3, we define the interpretation of our theory over the new model. In Section A.4, we prove that this interpretation over this model satisfies both the original physical theory and the axioms of knowledge and communication.

A.1. A physical theory

A physical theory is a set of constraints on actions and fluents. A communicative action may have physical preconditions, effects, or other constraints, but these may not depend on the *content* of the communication. That is, from the physical point of view, communicative actions are distinguished only by the identity of the speaker and hearer, not the content. Physical theories do not refer to knowledge states.

Our objective here is to prove that any reasonable physical theory is consistent with our theory of knowledge and communication. To do this, we have to ensure that the two theories "join up", so to speak; specifically, that the physical theory does not impose any constraints that are incompatible with the epistemic theory. There are three potential sources of trouble.

- Axioms I.1, I.2, and I.4 together imply that, if *AS* can communicate with *AH* then, in general, there are a large number of different possible communicative acts that *AS* can perform. Specifically, in any situation S, if $Q1$ and $Q2$ are fluents such that (a) *AS* knows that both $Q1$ and $Q2$ hold; but (b) it is not shared knowledge between *AS* and *AH* that $Q1 \Leftrightarrow Q2$, then the act of *AS* informing *AH* that $Q1$ different from the act of *AS* informing *AH* that $Q2$. The physical theory could make this impossible by asserting that only a small number of different communicative acts are feasible in S. For instance, the statement that only two different communicative acts are feasible in s0 could be stated in the formula

  $$\exists_{S1a,S1b}\ \text{occurs}(\text{do}(\text{as, communicate(ah)}), \text{s0}, S1a) \wedge$$
  $$\text{occurs}(\text{do}(\text{as, communicate(ah)}), \text{s0}, S1b) \wedge S1a \neq S1b \wedge$$
  $$\forall_{S1}\ \text{occurs}(\text{do}(\text{as, communicate(ah)}), \text{s0}, S1) \Rightarrow [S1 = S1a \vee S1 = S1b]$$

  To block this, we impose condition (4) in Definition A.6 below: a physical theory must be consistent with the constraint that, if any communicative action is feasible in a situation, then infinitely many physically indistinguishable actions are feasible in that situation.
- Axiom I.5 asserts the existence of a large number of fluents. The physical theory could assert that only a limited class of fluents exist. E.g., the following axiom asserts that the only fluents have the form "on(A, B)" where A and B are blocks.

  $$\forall_Q\ \exists_{A,B}\ \text{block}(A) \wedge \text{block}(B) \wedge Q = \text{on}(A, B)$$

  This is not at all far-fetched; one approach to the frame problem is to assert "The only fluents changed by action A are $Q1 \ldots Qk$", which leads to the same kind of problem.

We get around this problem by distinguishing between *physical fluents* and *general fluents*, and requiring that a physical theory can only refer to physical fluents.

- Similarly, the theory of communication requires the existence of actionals “inform(AH, Q)” and of actions “do(AS, inform(AH, Q)).” We have to make sure that the physical theory does not simply prohibit these; e.g. assert that the only possible actionals have the form “communicate(AH)” and “puton(A, B)”. To insure this, we require that the physical theory can only refer to physical actions and actionals.

Definition A.1. A *physical language* is a first-order language containing the sorts “situations”, “agents”, “physical actionals”, “physical actions”, “physical fluents”, and “clock times”; containing the non-logical symbols, “$<$”, “do”, “occurs”, “holds”, “time”, and “communicate”; and excluding the symbols, “k_acc”, “inform”, and “sk_acc”. (The language may or may not contain any sort or non-logical symbol other than those mentioned above.)

Definition A.2. Let $\mathcal{L}$ be a physical language. Let $\mathcal{M}$ be a model and let $\mathcal{I}$ be an interpretation of $\mathcal{L}$ in $\mathcal{M}$. Let s0 and s1 be situations in $\mathcal{M}$. Situation s1 is a *successor* of s0 if s0 $<$ s1 and there is no situation sm such that s0 $<$ sm $<$ s1.

Here, and in subsequent definitions, we implicitly use $\mathcal{I}$ to apply nomenclature from $\mathcal{L}$ to entities in $\mathcal{M}$. More formal statements of the condition “s0 $<$ s1” above would be, “The pair $\langle$s0, s1$\rangle \in \mathcal{I}$(‘$<$’)” or “The open formula $SA < SB$ is true in $\mathcal{M}$ under $\mathcal{I}$ under the valuation $SA \rightarrow$ s0, $SB \rightarrow$ s1”. We will use the shorter form when it is clear; when necessary, we will be more precise.

Definition A.3. Let $\mathcal{L}$, $\mathcal{M}$, $\mathcal{I}$ be as above. Let s0, s1 be situations in $\mathcal{M}$. We say that s1 is a *communication successor* of s0 if s1 is a successor of s0 and there exist agents as,ah and a situation sz such that s1 $\leqslant$ sz and occurs(do(as,communicate(ah)),s0,sz).

Definition A.4. Let $\mathcal{L}$, $\mathcal{M}$, $\mathcal{I}$ be as above. Let τ be a function from $\mathcal{M}$ to itself which is one-to-one and onto. The function τ is said to be a *situational automorphism* if the following conditions hold:

(1) If X is not a situation, then $\tau(X) = X$.
(2) Let α be a predicate symbol in $\mathcal{L}$ with k arguments or a function symbol with $k-1$ arguments. Note that, under standard Tarskian semantics, $\mathcal{I}(\alpha)$ is a set of k-tuples of elements of $\mathcal{M}$. A tuple $\langle x_1 \dots x_k\rangle \in \mathcal{I}(\alpha)$ if and only if $\langle \tau(x_1) \dots \tau(x_k)\rangle \in \mathcal{I}(\alpha)$.

Definition A.5. Two situations SA and SB are *indistinguishable* if the following holds: Let SSA be the part of the time structure following SA and SSB be the part of the time structure following SB.

$$SSA = \{S \in \mathcal{M} \mid SA \leqslant S\}$$
$$SSB = \{S \in \mathcal{M} \mid SB \leqslant S\}$$

Then there exists a situational automorphism τ over $\mathcal{M}$ such that $\tau(SSA) = SSB$, $\tau(SSB) = SSA$, and for any situation S which is not in *SSA* and *SSB*, $\tau(S) = S$.

Definition A.6. Let $\mathcal{L}$ be a physical language, and let $\mathcal{T}$ be a theory over $\mathcal{L}$. $\mathcal{T}$ is an *acceptable physical theory* (i.e., acceptable for our discussion here) if there exists a model $\mathcal{M}$ and an interpretation $\mathcal{I}$ of $\mathcal{L}$ over $\mathcal{M}$ such that the following conditions are satisfied:

(1) $\mathcal{I}$ maps the sort of clock times to the positive integers, and the relation $T1 < T2$ on clock times to the usual ordering on integers.
(2) $\mathcal{M}$ satisfies axioms T.1–T.9 in Table 1 under $\mathcal{T}$, where T.8 and T.9 are restricted to physical actions.
(3) $\mathcal{M}$ satisfies theory $\mathcal{T}$ under $\mathcal{I}$.
(4) For any situations s0, s1 and agents as, ah in $\mathcal{M}$, if s1 is a communication successor of s0, then there are infinitely many successors of s0 that are physically indistinguishable from s1.
(5) If α is a predicate symbol in $\mathcal{L}$ with more than one situational argument, then $\alpha(X_1 \ldots X_k)$ holds only if all the situations among $X_1 \ldots X_k$ are ordered with respect to $<$. (Note that this condition holds both when α is "$<$" and α is "occurs".) If $\beta(X_1 \ldots X_k)$ is a function symbol, then the above condition holds for the relation $X_{k+1} = \beta(X_1 \ldots X_k)$.

We can now state precisely the theorem that is the objective of this appendix.

Theorem 1. *Let $\mathcal{T}$ be an acceptable physical theory, and let $\mathcal{A}$ be $\mathcal{T}$ together with axioms* K.1–K.8 *and* I.1–I.5*, and with* T.8 *and* T.9 *extended to arbitrary actions. Then $\mathcal{A}$ is consistent.*

Sections A.2–A.4 give the proof of this theorem.

A.2. Model construction

Sketch of model construction

The main sticking point of the proof is as follows: In order to satisfy the comprehension axiom, we must define a fluent to be any set of situations. However, if Q is a fluent, then the act of *AS* informing *AH* of Q in $S1$ generates a new situation; and if we generate a separate "inform" act for each fluent, then we would have a unsolvable vicious circularity.

We are rescued here by axiom I.4 together with the theorem, proven in Theorem 3 below, that, in a discrete time structure satisfying the axiom of memory (K.4), knowledge accessibility relations can only connect situations of the same time, and therefore the current time is always common knowledge between all agents. Let q1 be any fluent that holds in situation s1. By axiom I.4, if *AS* informs *AH* of q1 over the interval [s1, s2] and *AS* and *AH* have shared knowledge that q1 $\Leftrightarrow$ q2 in s1, then the same act can be characterized as *AS* informing *AH* of q2. Let t1 = time(s1). Let q2 be the fluent such that holds(S, q2) $\Leftrightarrow$ holds(S, q1) $\wedge$ time(S) = t1. Then *AS* and *AH* have shared knowledge in s1 that q1 is equivalent to q2. Therefore, it suffices to generate an occurrence of an inform act starting

Table A.1
Construction of a model

```
Constructing a model
procedure model_construct(in 𝒯: an acceptable physical theory;
                               ℳ: a model of theory 𝒯)
                             return a structure of u-situations over which we will define
                                    a model of the extended theory.
for each p-situation PS in ℳ, construct a u-situation US.
    Label PHYS(US) = PS, time(US) = 0.
for (each agent A), define the relation K_ACC(A,·,·)
    to be some equivalence relation over the u-situations constructed above.
for (K = 0 to ∞) do {
  for (each u-situation S of time K) do {
    for (each p-situation PS following PHYS(S) in ℳ)
      construct a new u-situation S1 and mark PHYS(S1) = PS;
    for (each pair of agents AS, AH) do {
      if (in ℳ there is an act starting in S of AS communicating to AH)
      then {
        SSL := the set of u-situations knowledge-accessible from S
                  relative to the knowledge of AS;
        SSU := the set of u-situations knowledge-accessible from S
                  relative to the shared knowledge of AS and AH;
        for (each set SS that is a subset of SSU and a superset of SSL) do {
          construct an action "do(AS,inform(AH,SS))" starting in S;
          construct a successor S1 of S corresponding to the execution of this action;
          label PHYS(S1) to be a u-situation in ℳ following a communicate action in PHYS(S);
        }
      } } }
    use the axioms of knowledge to construct a valid set of
        knowledge accessibility relations over the new u-situations
} return (the set of u-situations plus the set of knowledge accessibility relations)
```

in $S1$ only for fluents like q2 that specify the current time, and such a fluent can be identified with a set of situations of the same time as $S1$. This limitation allow us to break the circularity in the construction of situations and informative acts: the content of informative acts starting at time K is a subset of the situations whose time is K; informative acts starting in time K generate situations whose time is $K+1$.

Therefore, we can use the "algorithm" shown in Table A.1 to construct a model of the theory $\mathcal{A}$. The main difference between the model $\mathcal{M}$ of theory $\mathcal{T}$ and the model $\mathcal{U}$ of $\mathcal{A}$ is that $\mathcal{U}$ contains many more situations. To avoid confusion, we will call the situations of $\mathcal{M}$ "p-situations" and call the situations of $\mathcal{U}$ "u-situations". Each u-situation US has a corresponding p-situation, denoted PHYS(US), which is physically indistinguishable from US. The difference is that US may associate specific contents with some of the communication actions that precede PHYS(US).

Theorem 3. *If the set of clocktimes is equal to the positive integers, then for any situations* SA, SB, *if* $\text{k\_acc}(A, SA, SB)$ *then* $\text{time}(SA) = \text{time}(SB)$.

Proof. Suppose that time(SA) $<$ time(SB) $= k$. By axioms T.7, T.6 and T.5, there exist situations $SB_0 < SB_1 < \cdots < SB_{k-1} < SB$ such that time(SB_i) $= i$. By axiom K.4 there exist $SA_0 \ldots SA_{k-1}$ such that k_acc(A, SA_i, SB_i), $SA_{i-1} < SA_i$ and $SA_{k-1} < SA$; but this is impossible, since time(SA) $< k$. □

Formal construction of the model

The definitions in this section essentially amount to a formalized re-statement of the "algorithm" in Table A.1.

Let $\mathcal{L}$ be a physical language. Let $\mathcal{T}$ be an acceptable physical theory over $\mathcal{L}$. Let $\mathcal{M}$ be a model and let $\mathcal{I}$ be an interpretation of $\mathcal{L}$ satisfying the conditions of Definition A.6.

The remaining definitions in this section are relative to a fixed choice of $\mathcal{L}$, $\mathcal{T}$, $\mathcal{M}$, and $\mathcal{I}$.

For convenience, for each symbol τ in $\mathcal{T}$, including sorts, we use the same symbol in block capitals to denote the image of τ under $\mathcal{I}$; this is an individual, a subset, a mapping, or a relation over $\mathcal{M}$. Thus, for example, AGENTS is the image under $\mathcal{I}$ of the sort "agents"; TIME is the image under $\mathcal{I}$ of the function symbol "time" and so on.

We now proceed to building up the set of u-situations. This construction is recursive over time. Naturally, the base case is at time 0.

The most important and complex part of the construction is the wider class of situations that we will need. In general a u-situation US is a pair ⟨S1, MM⟩ where:

- S1 is a p-situation. We will write S1 = PHYS(US).
- MM is a set of 4-tuples ⟨AS, AH, USSQ, SX⟩. AS and AH are agents; USSQ is a set of u-situations; and SX is a p-situation such that SX $<$ S1 and such that OCCURS(DO(AS, COMMUNICATE(AH)), SX, SZ), for some SZ that is ordered with respect to S1. Such a tuple asserts that an action of AS informing AH of content USSQ began in a u-situation USX $<$ US. We write MM = MM(US).

It will be convenient to posit the existence of an atomic entity **INFORM**, which is not in $\mathcal{M}$, and of an entity **DO**.

Definition A.7. Let PS be a p-situation such that TIME(PS) = 0. A *u-situation at time* 0 is a pair of the form US = $\langle \text{PS}, \emptyset \rangle$. The function ANCESTOR(US) maps a u-situation US to a set of u-situations, the ancestors of US in the time structure.

Definition A.8. A *time structure of depth* 0 TS is a pair:

- The set of u-situations U_SITS = $\{\langle \text{PS}, \emptyset \rangle \mid \text{PS} \in \text{SITUATIONS},\ \text{TIME}(\text{PS}) = 0\}$ with one u-situation for each p-situation at time 0.
- A function K_ACC mapping any agent A ∈ AGENTS to an equivalence relation over U_SITS.

Definitions A.9 through A.15 are mutually recursive over the depth k, successively building up the model forward in time.

Definition A.9. Let TS be a time structure of depth K. Let US be a u-situation of time K in TS. Let S1 = PHYS(US). Let MM be a collection of 4-tuples as described above. Let S2 be a successor to S1. The *simple successor to* US *parallel to* S2 is the pair $\langle$S2,MM$\rangle$.

Definition A.10. Let TS = $\langle$U_SITS, K_ACC$\rangle$, US, S1, MM be as above. Let AS and AH be agents. A *possible communicative content* from AS to AH is a set of u-situations USSQ of time K in U_SITS satisfying the following: let USSL be the set of u-situations USA in TS such that $\langle$US1, USA$\rangle \in$ K_ACC(AS). Let USSU be the set of u-situations USA in USSL such that there exist $\mathrm{US}_0 = \mathrm{US}$, US_1, $\mathrm{US}_2, \ldots, \mathrm{US}_N = \mathrm{USA}$, such that for each J, $\langle \mathrm{US}_J, \mathrm{US}_{J+1} \rangle$ is either in K_ ACC(AS) or in K_ACC(AH). Then USSL $\subseteq$ USSQ $\subseteq$ USSU.

The 4-tuple $\langle$AS, AH, USSQ, S1$\rangle$ is called an *inform indicator* starting in S1.

Definition A.11. Let TS, US, S1, MM be as above. Let S2 be a successor of S1. Let I = $\langle$AS, AH, USSQ, S1$\rangle$ be an inform indicator starting in S1. I *possibly leads toward* S2 if there exists SZ $\geqslant$ S2 such that OCCURS(DO(AS, COMMUNICATE(AH)), S1, SZ). An *informative sheaf* in US toward S2 is a set MMX of inform indicators in US toward S2 such that no two elements of MMX have the same speaker and the same hearer. An *informative successor to* US toward S2 is a pair $\langle$S2, MM2$\rangle$ where MM2 is the union of MM with some informative sheaf in US toward S2.

Definition A.12. Let TS, US, S1, S2 be as above. A *u-successor set* for US toward S2 is the union of

- The simple successor to US, S2.
- A set USS of informative successors to US, S2 with the following property: if M is any inform indicator in S1, then there exists an element $\langle$S2, MM$\rangle \in$ USS such that M $\in$ MM. That is, every inform indicator is attached to at least one successor of US.

A u-successor of a u-situation at time K is a u-situation at time $K+1$. If US1 is a u-successor of US then ANCESTOR(US1) = ANCESTORS(US) $\cup$ {US}.

Definition A.13. Let TS be a time-structure of depth K. A *u-situation successor space* for TS is the union over [all u-situations US of depth K in TS] and [all successors S2 of PHYS(US))] of some u-successor set for US, S2.

Definition A.14. Let TS = $\langle$U_SITS, K_ACC$\rangle$ be a time-structure of depth K. Let USA and USB be u-situations of depth K in TS. Let US1A be a u-successor of USA and let US1B be a u-successor of USB. Let A be an agent. Then US1B is *possibly knowledge accessible* from US1A relative to A if all the following conditions hold:

- $\langle$USA, USB$\rangle \in$ K_ACC(A).
- For any actional Z and p-situations SXA, SYA, if OCCURS(DO(A, Z), SXA, SYA) and SXA $\leqslant$ PHYS(USA), then

 – If SYA < PHYS(USA), then there exist SXB, SYB such that OCCURS(DO(A, Z), SXB, SYB) and SYB < PHYS(USB).
 – If SYA = PHYS(USA), then there exists SXB such that OCCURS(DO(A, Z), SXB, PHYS(USB)).
 – If SXA < PHYS(USA) < SYA, then there exist SXB, SYB such that OCCURS(DO(A, Z), SXB, SYB) and SXB < PHYS(USB) < SYB.
 – If SXA = PHYS(USA) < SYA, then there exists SYB such that OCCURS(DO(A, Z), PHYS(USB), SYB).
- If there exists a tuple ⟨AS, A, USSQ, SX⟩ in MM(USA) and OCCURS(DO(AS, COMMUNICATE(AH)), SX, PHYS(USA)) then there exists a p-situation SXB and a tuple ⟨AS, A, USSQ, SXB⟩ in MM(USB) and OCCURS(DO(AS, COMMUNICATE(AH)), SXB, PHYS(USB)). (That is, if AS has completed informing A of USSQ, then A knows that AS has completed informing him of USSQ.)

Definition A.15. Let TS be a time-structure of depth K. A *possible successor* to TS is a pair TS1 = ⟨U_SITS1, K_ACC1⟩ where

- U_SITS1 is a u-situation successor space for TS;
- for each agent A ∈ AGENTS, K_ACC1(A) is an equivalence relation over U_SITS1, which is a subset of the relation, “USB is possibly knowledge accessible from USA”. (Note that, since all the conditions on “possibly knowledge accessible” have the form “Some property holds on US1A iff the corresponding property holds on US1B”, the relation “possibly knowledge accessible relative to (A)” is itself always an equivalence relation.)

TS1 is said to be of depth K + 1.

Finally, we let this construction go from time 0 to infinity.

Definition A.16. Let $\mathrm{TS}_0 = \langle \mathrm{U\_SITS}_0, \mathrm{K\_ACC}_0\rangle$, $\mathrm{TS}_1 = \langle \mathrm{U\_SITS}_1, \mathrm{K\_ACC}_1\rangle$, ... be a sequence such that TS_0 is a time structure of depth 0 and for each i, TS_{i+1} is a possible successor for TS_i. Then the pair

$$\mathrm{TS}_\infty = \langle \mathrm{U\_SITS}_\infty, \mathrm{K\_\ ACC}_\infty\rangle = \left\langle \bigcup_i \mathrm{U\_SITS}_i, \bigcup_j \mathrm{K\_\ ACC}_j \right\rangle$$

is a *communicative model extension of* $\mathcal{M}, I$.

A.3. Interpretation

Let $\mathcal{L}$, $\mathcal{M}$ and $\mathcal{I}$ be as in the previous section. Let $\mathcal{W}$ be the language $\mathcal{L}$ combined with the following additional elements:

- The sorts “fluent”, “actional” and “actions”, which are super-categories of the sort “physical fluent”, “physical action”, and “physical actional”, respectively.

- The symbols “k_acc”, “sk_acc”, and “inform”.

Let $\mathrm{TS}_\infty = \langle \mathrm{U\_SITS}_\infty, \mathrm{K\_ ACC}_\infty \rangle$ be a communicative model extension of $\mathcal{M}, I$.

In this section, we define an interpretation $\mathcal{J}$ of $\mathcal{W}$ in terms of constructions over TS_∞ and $\mathcal{M}$. For notational convenience, we will write the image of a symbol under $\mathcal{J}$ by writing it in lower-case boldface; thus, for example, **sk_acc** $= \mathcal{J}$(“sk_acc”). We will use ordinary Roman font where symbols are used in prefix notation and are interpreted under $\mathcal{J}$. For example, if we write “occurs($E, S1, S2$)” we mean the interpretation of “occurs” under $\mathcal{J}$. Note that, if a symbol is in $\mathcal{L}$, then its interpretation under $\mathcal{I}$ may be different than its interpretation under $\mathcal{J}$.

We will first discuss the construction of $\mathcal{J}$ informally and then proceed to the formal definition.

The first issue is fluents. On the one hand, axiom I.5 asserts that every property of situations $\alpha(S)$ has an associated fluent Q_α such that Q_α holds in just those situations satisfying S. The usual extensionalizing trick, therefore, is to identify Q_α with the set of u-situations satisfying α; generally, to identify fluents with sets of situations. On the other hand, to extend the theory $\mathcal{T}$ to the new model, we must make sure that every physical fluent in $\mathcal{T}$ is still a fluent in the new theory. Moreover it is possible that $\mathcal{T}$ involves the existence of two different fluents that are in fact coextensional in terms of the situations where they hold, but differ in terms of some other property of interest to $\mathcal{T}$. Therefore, we define a general fluent as a pair of a label and a set of u-situations. For a physical fluent that is, so to speak, grandfathered from $\mathcal{T}$, the label is just the physical fluent; for all other fluents, the label is immaterial. A physical fluent Q holds in u-situation S just if Q holds in PHYS(S).

The second issue is the occurrence of actions. For physical actions, as for physical fluents, we use the “PHYS” mapping to guide us; a physical action E occurs from US1 to US2 if E occurs from PHYS(US1) to PHYS(US2). For informative events, there are two steps. First, axiom I.4 asserts that “do(as, inform(ah, q1))” and “do(as, inform(ah, q2))” co-occur from us1 to us2 if the intersection of q1 with the set of u-situations that are sk-accessible relative to as,ah from us1 is the same as the intersection of us2 with that set. Second, the occurrence from us1 to us2 of the act “do(as, inform(ah, q0))” where q0 is a subset of the sk-accessible situations is indicated in the second (MM) field of the u-situation us1.

Finally for simplicity we assume that there are no “pointless coincidences” between $\mathcal{M}$ and the constructions we will use in $\mathcal{J}$. That is to say: It is conceivable that $\mathcal{M}$ itself happens to contain, as an entity, some tuple that we will want to define as an entity in the denotation of $\mathcal{J}$. Such a coincidence would cause propositions to be true and false in ways that we do not intend. One could block this by modifying Definition A.20 below as follows: Wherever the definition constructs an tuple, add an additional element that is not an element of $\mathcal{M}$ (e.g., $\mathcal{M}$ itself). That will block any such coincidences. For the sake of readability, I have omitted these.

Otherwise, the definition is pretty much straightforward.

Definition A.17. A *general fluent* is a pair $\langle \mathrm{LABEL}, \mathrm{USS} \rangle$ where LABEL is either a physical fluent or 0, and USS is a set of u-situations.

Definition A.18. For any PF in PHYSICAL-FLUENTS, define PF_MAP(PF) to be the pair $\langle$PF, {US | US $\in$ U-SITUATIONS $\wedge$ HOLDS(PHYS(US), PF)}$\rangle$.

Define PF_IMAGES = {PF_MAP(PF) | PF $\in$ PHYSICAL-FLUENTS}.

Definition A.19. We define a general mapping "U2P_MAP" from constructions over TS_∞ to entities in $\mathcal{M}$ as follows:

- If U is a u-situation, then U2P_MAP(U) = PHYS(U).
- If U = $\langle$PF,USS$\rangle$ $\in$ PF_IMAGES then U2P_MAP(U) = PF.
- If U $\in \mathcal{M}$ then U2P_MAP(U) = U.
- Else U2P_MAP(U) is undefined.

In reading Definition A.20 below, keep in mind that, in the standard Tarskian semantics for first-order logic, the denotation of a function or a predicate symbol is a set of tuples. Similarly, we take the denotation of a sort to be a set of entities.

Definition A.20 (*Long*). Let $\mathcal{L}$, M, I, W, U be as above. We define the function $\mathcal{J}$ over the sorts and symbols of $\mathcal{W}$ as follows:

Sorts:

$\mathcal{J}$(the sort "clock time") = the non-negative integers.
$\mathcal{J}$(the sort "agent") = $\mathcal{I}$("agent").
$\mathcal{J}$(the sort "situation") = the set of u-situations in $\mathcal{U}$.
$\mathcal{J}$(the sort "fluent") = the set of general fluents.
$\mathcal{J}$(the sort "physical fluent") = PF_IMAGES.
$\mathcal{J}$(the sort "physical actional") = $\mathcal{I}$("physical actional").
$\mathcal{J}$(the sort "physical action') = $\mathcal{I}$("physical action").
Let **informative_actionals** $\equiv$ {$\langle$**INFORM**, AH, Q$\rangle$ | AH $\in$ **agent** $\wedge$ Q $\in$ **fluent**}.
Let **informative_actions** $\equiv$ {$\langle$**DO**, A, Z$\rangle$ | Z $\in$ **informative_ actionals**}.
$\mathcal{J}$(the sort "actional") = $\mathcal{I}$("physical actional") $\cup$ **informative_actionals**.
$\mathcal{J}$(the sort "action") = $\mathcal{I}$("physical action") $\cup$ **informative_actions**.
If σ is any other sort used in $\mathcal{L}$, then $\mathcal{J}(\sigma) = \mathcal{I}(\sigma)$.

Non-logical symbols:

$\mathcal{J}$("<") (as a predicate on clock times) = the usual ordering on integers.
$\mathcal{J}$("<") (as a predicate on situations) = {$\langle S1, S2\rangle$ | $S1, S2 \in$ **situation** and $S1$ is an ancestor of $S2$}.
$\mathcal{J}$("holds") = {$\langle S, Q\rangle$ | $S \in$ **situation**, $Q = \langle PF, USS\rangle \in$ **fluent** and $S \in USS$}.
$\mathcal{J}$("time") = {$\langle S, T\rangle$ | $S \in$ **situation**, $T \in$ **clock time** and S is of time T}.
$\mathcal{J}$("communicate") = $\mathcal{I}$("communicate").
$\mathcal{J}$("do") = $\mathcal{I}$("do") $\cup$ {$\langle$A, Z, $\langle$**DO**, A, Z$\rangle\rangle$ | A $\in$ **agent** and Z $\in$ **informative_actionals**}.
$\mathcal{J}$("inform") = {$\langle$AH, Q, $\langle$**INFORM**, AH, Q$\rangle\rangle$ | AH $\in$ **agent** and Q $\in$ **fluent**}.
$\mathcal{J}$("k_acc") = {$\langle$A, S1, S2$\rangle$ | A $\in$ **agents** and $\langle$S1, S2$\rangle$ $\in$ $\mathrm{K\_ACC}_\infty$(A)}.

$\mathcal{J}$(“sk_acc”) =
{⟨AS, AH, SA, SB⟩ |
exists($S_0 = SA$, $S_1 \ldots S_k = SB$) such that
for $(i = 1 \ldots k)$ either **k_acc**(AS, S_{i-1}, S_i) or **k_acc**(AH, S_{i-1}, S_i)
}.

$\mathcal{J}$(“occurs”) =
{⟨E,US1,US2⟩ |
E ∈ **action** and US1, US2 ∈ **situation** and $US1 < US2$ and
either [E ∈ $\mathcal{I}$(“physical action”) and OCCURS(E, PHYS(US1), PHYS(US2))] or
[there exist (A,AH ∈ **agent**; Q1, Q2 ∈ **fluent**; USS1, USS2) such that
E = ⟨**DO**, AS, ⟨**INFORM**, AH, Q1⟩⟩ and
Q1 = ⟨PF1, USS1⟩, Q2 = ⟨PF2, USS2⟩;
USS2 = {US ∈ USS1 | ⟨AS, AH, US1, US⟩ ∈ **sk_acc**},
OCCURS(DO(AS, COMMUNICATE(AH)), PHYS(US1), PHYS(US2)) and
⟨AS, AH, USS2, PHYS(US1)⟩ ∈ MM(US2)
]
}.

Let α be any symbol in $\mathcal{L}$ other than those enumerated above. $\mathcal{I}(\alpha)$ is a set of tuples of entities in $\mathcal{M}$. A tuple T' is a replacement for tuple T if, for each index I, U2P_MAP($T'[I]$) = $T[I]$. Then $\mathcal{J}(\alpha)$ is the set of all replacements R for the tuples in $\mathcal{I}(\alpha)$, such that any two situations in R are ordered under $\mathcal{J}$(“<”).

Definition A.21. The model $\mathcal{U}$ is the union of **clocktime**, **agent**, **situation**, **fluent**, **actional**, **action** and $\mathcal{M}$.

Note that the function U2P_MAP(X) is defined for exactly those entities X which are in $\mathcal{J}(\sigma)$ where σ is one of the sorts in the physical language (clock times, situations, agents, physical fluents, physical actionals, physical actions, and other sorts in $\mathcal{L}$).

A.4. Soundness

Throughout this section: Let $\mathcal{L}$ be a physical language. Let $\mathcal{T}$ be an acceptable physical theory over $\mathcal{L}$. Let $\mathcal{M}$ be a model and let $\mathcal{I}$ be an interpretation of $\mathcal{L}$ in $\mathcal{M}$ that satisfies $\mathcal{T}$. Let $\mathcal{U}$ and $\mathcal{J}$ be defined as above.

We will assume that $\mathcal{L}$ is strongly sorted; in particular, that every variable in $\mathcal{L}$ is labelled with its sort. A valuation over variables in $\mathcal{L}$ is required to respect the sort constraint. That is, if μ_i is a variable of sort σ_i, and V is a valuation of μ_i in $\mathcal{M}$ then $\mathrm{V}(\mu_i) \in \mathcal{I}(\sigma_i)$. If W is a valuation of μ_i in $\mathcal{U}$ then $\mathrm{W}(\mu_i) \in \mathcal{J}(\sigma_i)$.

Lemma A.1. *For every p-situation* PS *in* $\mathcal{M}$ *there exists a u-situation* US *in* $\mathcal{U}$ *such that* PHYS(US) = PS.

Proof. By induction on TIME(PS). If TIME(PS) = 0 then there exists a corresponding u-situation by Definition A.8. Suppose the statement is true for all PS such that

TIME(PS) $= k$. Let PS1 be a p-situation such that TIME(PS1) $= k + 1$. By axiom T.7, PS1 is the successor of some situation PS0 such that TIME(PS0) $= k$. By the induction hypothesis, there is a situation US0 such that PHYS(US0) = PS1. By Definition A.9 there is a simple successor US1 of US0 such that PHYS(US1) = PS1. □

Lemma A.2. *For any u-situation U,* TIME(PHYS(U)) = TIME(U). *For any two u-situations $U1, U2$ if $U1 < U2$ then* PHYS($U1$) < PHYS(US2).

Proof. Immediate from the definition of $\mathcal{J}$("<") in Definition A.20 and the definition of "ANCESTORS" in Definitions A.7 and A.12. □

Lemma A.3. *Let $\mu_1 \ldots \mu_k$ be variables in $\mathcal{L}$. Let* V *be a valuation mapping each variable μ_i into $\mathcal{I}(\sigma_i)$. Then there exists a valuation* W *into $\mathcal{U}$ such that* U2P_MAP(W(μ_i)) = V(μ_i).

Proof. Immediate from Lemma A.1 together with the construction of U2P_MAP and the fact that, for each sort σ, U2P_MAP maps an element of $\mathcal{J}(\sigma)$ to an element of $\mathcal{I}(\sigma)$. □

Lemma A.4. *Let $\alpha(\mu_1 \ldots \mu_k)$ be a predicate symbol in $\mathcal{L}$, including equality, where $\mu_1 \ldots \mu_k$ have sorts in $\mathcal{L}$. Let* W *be a valuation from μ_i into $\mathcal{U}$. Define* V(μ_i) = U2P_MAP(W(μ_i)). *Then $\alpha(\mu_1 \ldots \mu_k)$ holds in $\mathcal{U}$ under $\mathcal{J}$,* W *if and only if* (a) *$\alpha(\mu_1 \ldots \mu_k)$ holds in $\mathcal{M}$ under $\mathcal{I}$,* V *and* (b) *any two situations* W(μ_i) *and* W(μ_j) *are ordered under $\mathcal{J}$("<").*

Proof. We must consider separately the cases where α is (A) equality over non-situations; (B) equality over situations; (C) the symbol "<" over clock times; (D) the symbol "<" over situations; (E) the symbol "occurs"; (F) the symbol "holds"; (G) any other predicate symbol in $\mathcal{L}$.

(A) Equality over non-situations: from Definitions A.19 and A.20.
(B) Equality over situations: following Definitions A.19 and A.20, this amounts to the claim that US1 = US2 if and only if PHYS(US1) = PHYS(US2) and US1 and US2 are ordered. The implication from left to right is trivial. For the implication from right to left, consider that, if US1 and US2 are ordered but US1 ≠ US2, then either US1 < US2 or US2 < US1. If US1 < US2, then time(US1) < time(US2) so by Lemma A.2, PHYS(US1) ≠ PHYS(US2); and likewise if US2 < US1.
(C) The symbol "<" over clock times: from the fact that the interpretation is the same under $\mathcal{J}$ as under $\mathcal{I}$ (Definition A.20).
(D) The symbol "<" over situations: analogous to (B) above.
(E) The symbol "occurs". By Definition A.20, if E is a physical action then occurs(E, $S1$, $S2$) occurs under $\mathcal{J}$ if and only if occurs(E, PHYS($S1$), PHYS($S2$)) under $\mathcal{I}$.
(F) Let μ_1, μ_2 be variables of sorts "situation" and "physical fluent" respectively. Let PF = V(μ_1). Since U2P_MAP(W(μ_2)) = V(μ_2) = PF, by Definition A.19 W(μ_2) ∈ PF_IMAGES, which, by Definitions A.18 and A.19, means that W(μ_2) = ⟨PF, {US ∈

U-SITUATIONS | HOLDS(PHYS(US), PF)}⟩ By Definition A.20 it follows that $\langle \mathrm{W}(\mu_1), \mathrm{W}(\mu_2)\rangle \in \mathcal{J}$(“holds”) if and only if $\langle \mathrm{V}(\mu_1), \mathrm{PF}\rangle \in \mathcal{I}$(“holds”).

(G) α is any other predicate symbol in $\mathcal{L}$. Immediate from Definition A.20. □

Lemma A.5. *Let $\beta(\mu_1 \ldots \mu_k)$ be a function symbol in $\mathcal{L}$, where $\mu_1 \ldots \mu_k$ have sorts in $\mathcal{L}$. Let* W *be a valuation from μ_i into $\mathcal{U}$ such that, for any two situational variables μ_p and μ_q,* $\mathrm{W}(\mu_p)$ *and* $\mathrm{W}(\mu_q)$ *are ordered with respect to $\mathcal{J}$(“<”). Define* $\mathrm{V}(\mu_i) = \mathrm{U2P\_MAP}(\mathrm{W}(\mu_i))$. *Then the value of $\beta(\mu_1 \ldots \mu_k)$ in $\mathcal{M}$ under $\mathcal{I}$,* V *is the image under* U2P_MAP *of the value of $\beta(\mu_1 \ldots \mu_k)$ in $\mathcal{U}$ under $\mathcal{J}$,* W.

Proof. As in the proof of Lemma A.4, we must consider separately the cases where β is (A) the function symbol “do”; (B) the function symbol “time”; (C) any other function symbol in $\mathcal{L}$.

(A) By Definitions A.19 and A.20, if A is an agent and Z is a physical actional then U2P_MAP($\mathcal{J}$(do(A, Z))) = $\mathcal{J}$(do(A, Z)) = $\mathcal{I}$(do(A, Z)) = $\mathcal{I}$(do(U2P_MAP(A), U2P_MAP(Z))). (Again, we are mildly abusing notation.)

(B) By Definitions A.19 and A.20, if US is a u-situation then U2P_MAP($\mathcal{J}$(time(US))) = $\mathcal{J}$(time(US)) = $\mathcal{I}$(time(PHYS(US))) = $\mathcal{I}$(time(U2P_MAP(US))).

(C) Let β be any other function symbol. Let $\langle x_1 \ldots x_k, y\rangle$ be any tuple where the x_i and y are entities in the image under $\mathcal{J}$ of the sorts in $\mathcal{L}$. Then by the last part of Definition A.20, $\langle x_1 \ldots x_k, y\rangle \in \mathcal{J}(\beta)$ iff $\langle \mathrm{U2P\_MAP}(x_1) \ldots \mathrm{U2P\_MAP}(x_k), \mathrm{U2P\_MAP}(y)\rangle \in \mathcal{I}(\beta)$.

But for any terms $\gamma_1 \ldots \gamma_k$ and any valuation W from the variables in the γ’s to $\mathcal{U}$, the denotation of $\beta(\gamma_1 \ldots \gamma_k)$ under $\mathcal{J}$, W is equal to y just if the tuple $\langle \mathcal{J}(\gamma_1) \ldots \mathcal{J}(\gamma_k), y\rangle$ is in $\mathcal{J}(\beta)$; and likewise for $\mathcal{I}$.

Unfortunately, U2P_MAP does not preserve truth-values of predicates over unordered u-situations; it is possible that U2P_MAP(US1) = U2P_MAP(US2) even though US1 ≠ US2, or that U2P_MAP(US1) < U2P_MAP(US2) even if US1 and US2 are unordered. There is, moreover, in general no way to modify U2P_MAP to preserve inequality, since the cardinality of the set of u-situations may be larger than the cardinality of p-situations. Therefore, in establishing below that if an open formula with inequalities or orderings is satisfiable in $\mathcal{J}$ then it is also satisfiable in $\mathcal{I}$, it is necessary to continuously “patch” the mapping U2P_MAP by mapping a u-situation US into some p-situation that is physically indistinguishable from U2P_MAP(US). Fortunately, we had the foresight to provide ourselves with plenty of these. Stating this exactly is a little involved; Definitions A.22–A.24 and Corollary A.6 through Lemma A.9 accomplish this.

Definition A.22. Let τ be a function from $\mathcal{U}$ to itself which is one-to-one and onto. The function τ is said to be a *physical automorphism* over $\mathcal{U}$ if the following conditions hold:

(1) If X is not a u-situation, then $\tau(X) = X$.

(2) Let $\alpha(\mu_1 \ldots \mu_k)$ be any atomic formula in $\mathcal{L}$ with free variables $\mu_1 \ldots \mu_k$. Let W and Y be valuations from μ_i to $\mathcal{U}$ such that $\mathrm{Y}(\mu_i) = \tau(\mathrm{W}(\mu_i))$. Then Y satisfies α only if W satisfies α.

Note that condition (2) only applies to formulas in the *physical* language $\mathcal{L}$, not in the broader language.

Definition A.23. Let S1, S2 be either two p-situations or two u-situations. Situation S is the *latest common ancestor* (LCA) of S1 and S2, if S $\leqslant$ S1, S $\leqslant$ S2 and S is the latest situation with that property. Since the order relation on situations is a forest of trees, any two situations have at most one latest common ancestor.

Definition A.24. Let $\langle \mu_1 \ldots \mu_k \rangle$ be a k-tuple of variables. Let W be a valuation of the μ's to $\mathcal{U}$ and let V be a valuation of the μ's to $\mathcal{M}$. V is said to be an *image* of W if the following conditions hold:

- If μ is not a situational variable, then $\mathrm{V}(\mu) = \mathrm{U2P\_MAP}(\mathrm{W}(\mu))$.
- There exists a physical automorphism τ over $\mathcal{U}$ such that, for each pair of situational variables μ_i, μ_j, if S is the latest common ancestor of $\mathrm{W}(\mu_i)$, $\mathrm{W}(\mu_j)$ then $\mathrm{PHYS}(\tau(\mathrm{S}))$ is the LCA of $\mathrm{V}(\mu_i)$, $\mathrm{V}(\mu_j)$; and if $\mathrm{W}(\mu_i)$ and $\mathrm{W}(\mu_j)$ have no common ancestor, then $\mathrm{V}(\mu_i)$ and $\mathrm{V}(\mu_j)$ have no common ancestor.

We say that the automorphism τ *establishes the correspondence* between W and V.

Corollary A.6. *Let* $\mu_1 \ldots \mu_k$, W, V, *and* τ *be as in Definition* A.24. *For each* i, $\mathrm{V}(\mu_i) = \mathrm{U2P\_MAP}(\tau(\mathrm{W}(\mu_i)))$.

If μ_i is a situational variable, then applying Definition A.24 and choosing $j = i$, since $\mathrm{W}(\mu_i)$ is the LCA of $\mathrm{W}(\mu_i)$ and itself, we have $\mathrm{U2P\_MAP}(\tau(\mathrm{W}(\mu_i))) = \mathrm{PHYS}(\tau(\mathrm{W}(\mu_i))) = \mathrm{LCA}(\mathrm{V}(\mu_i), \mathrm{V}(\mu_i)) = \mathrm{V}(\mu_i)$. If μ_i is not a situational variable, then the result is immediate.

Lemma A.7. *Let* μ_1 *and* μ_2 *be situational variables in* $\mathcal{L}$. *Let* W *and* V *be valuations of* μ_1, μ_2 *to* $\mathcal{U}$ *and* $\mathcal{M}$ *respectively, and let* V *be an image of* W. *Then* $\mathrm{W}(\mu_1) = \mathrm{W}(\mu_2)$ *if and only if* $\mathrm{V}(\mu_1) = \mathrm{V}(\mu_2)$ *and* $\mathrm{W}(\mu_1) < \mathrm{W}(\mu_2)$ *if and only if* $\mathrm{V}(\mu_1) < \mathrm{V}(\mu_2)$.

Proof. Let τ be an automorphism that establishes the correspondence between W and V. If $\mathrm{W}(\mu_1) = \mathrm{W}(\mu_2)$ then $\mathrm{V}(\mu_1) = \mathrm{V}(\mu_2)$, since $\mathrm{V}(\mu) = \mathrm{PHYS}(\tau(\mathrm{W}(\mu)))$ and is thus a function of $\mathrm{W}(\mu)$. If $\mathrm{W}(\mu_1) < \mathrm{W}(\mu_2)$ then by Lemma A.2, $\mathrm{V}(\mu_1) < \mathrm{V}(\mu_2)$.

Suppose that $\mathrm{V}(\mu_1) = \mathrm{V}(\mu_2)$. Thus, $\mathrm{LCA}(\mathrm{V}(\mu_1), \mathrm{V}(\mu_2)) = \mathrm{V}(\mu_1) = \mathrm{V}(\mu_2)$. By Definition A.24 $\mathrm{LCA}(\mathrm{W}(\mu_1), \mathrm{W}(\mu_2)) = \mathrm{W}(\mu_1) = \mathrm{W}(\mu_2)$.

Suppose that $\mathrm{V}(\mu_1) < \mathrm{V}(\mu_2)$. Thus, $\mathrm{LCA}(\mathrm{V}(\mu_1), \mathrm{V}(\mu_2)) = \mathrm{V}(\mu_1)$. By Definition A.24, $\mathrm{LCA}(\mathrm{W}(\mu_1), \mathrm{W}(\mu_2)) = \mathrm{W}(\mu_1)$. Therefore $\mathrm{W}(\mu_1) \leqslant \mathrm{W}(\mu_2)$. Since $\mathrm{V}(\mu_1) \neq \mathrm{V}(\mu_2)$, it follows from the earlier part of this lemma that $\mathrm{W}(\mu_1) \neq \mathrm{W}(\mu_2)$; hence $\mathrm{W}(\mu_1) < \mathrm{W}(\mu_2)$. $\square$

Lemma A.8. *Let $\alpha(\mu_1 \dots \mu_k)$ be a predicate symbol in $\mathcal{L}$. Let* W *be a valuation of the μ's to $\mathcal{U}$ and let* V *be an image of* W. *Then α holds in $\mathcal{U}$ under W if and only if α holds in $\mathcal{M}$ under* V.

Proof. Let τ be an automorphism that establishes the correspondence between W and V. Let $\mathrm{Q}(\mu_i) = \tau(\mathrm{W}(\mu_i))$. By Definition A.22, $\alpha(\mu_1 \dots \mu_k)$ holds under $\mathcal{J}$, W if and only if it holds under $\mathcal{J}$, Q. By Lemma A.4, $\alpha(\mu_1 \dots \mu_k)$ holds under $\mathcal{J}$, Q if and only if it holds under $\mathcal{I}$, V and for any two situational variables μ_a, μ_b, $\mathrm{Q}(\mu_a)$ and $\mathrm{Q}(\mu_b)$ are ordered. By Lemma A.7, $\mathrm{Q}(\mu_a)$ and $\mathrm{Q}(\mu_b)$ are ordered if and only if $\mathrm{V}(\mu_a)$ and $\mathrm{V}(\mu_b)$ are ordered; and by condition (5) of Definition A.6, $\alpha(\mu_1 \dots \mu_k)$ holds under $\mathcal{I}$, V only if $\mathrm{V}(\mu_a)$ and $\mathrm{V}(\mu_b)$ are ordered. Putting these together, it follows that $\alpha(\mu_1 \dots \mu_k)$ holds under $\mathcal{J}$, W if and only if it holds under $\mathcal{I}$, V. □

Lemma A.9. *Let $\beta(\mu_1 \dots \mu_k)$ be a function symbol in $\mathcal{L}$, and let μ_{k+1} be another variable. Let* W *be a valuation of the μ's to $\mathcal{U}$ and let* V *be an image of* W. *Then the equation $\mu_{k+1} = \beta(\mu_1 \dots \mu_k)$ holds in $\mathcal{U}$ under* W *if and only if it holds in $\mathcal{M}$ under* V.

Proof. Exactly analogous to the proof of Lemma A.8, substituting Lemma A.5 for Lemma A.4. □

Lemma A.10. *Let $\alpha(\mu_1 \dots \mu_k)$ be a quantifier-free formula in $\mathcal{L}$. Let* W *be a valuation of the μ's to $\mathcal{U}$ and let* V *be an image of* W. *Then α holds in $\mathcal{U}$ under* W *if and only if α holds in $\mathcal{M}$ under* V.

Proof. Straightforward structural induction over the form of α, using Lemmas A.8 and A.9. □

Lemma A.11. *Let $\mu_1 \dots \mu_k$ be variables whose sorts are in $\mathcal{L}$. Let* W *be a valuation from variables $\mu_1 \dots \mu_k$ to $\mathcal{U}$ and let* V *be an image of* W. (*We will include here the case where* $k = 0$; *in that case,* W *and* V *are null valuations.*) *Let μ_{k+1} be a new variable of sort σ_{k+1}.*

(1) *Let* A *be an entity in $\mathcal{J}(\sigma_{k+1})$. Let* $\mathrm{W}' = \mathrm{W} \cup \{\mu_{k+1} \to \mathrm{A}\}$. *Then there exists* B *in $\mathcal{M}$ such that* $\mathrm{V}' = \mathrm{V} \cup \{\mu_{k+1} \to \mathrm{B}\}$ *is an image of* W'.
(2) *Let* B *be an entity in $\mathcal{I}(\sigma_{k+1})$. Let* $\mathrm{V}' = \mathrm{V} \cup \{\mu_{k+1} \to \mathrm{B}\}$. *Then there exists* A *in $\mathcal{U}$ such that* V' *is an image of* $\mathrm{W}' = \mathrm{W} \cup \{\mu_{k+1} \to \mathrm{A}\}$.

Proof. Let τ be a physical automorphism over $\mathcal{U}$ that establishes the correspondence of W and V. If the sort of μ_{k+1} is not a situation, then both (1) and (2) are trivial; one can take A = B, leave the automorphism τ unchanged, and the result is immediate from the definitions. Therefore, we may assume that the sort of μ_{k+1} is a situation, and therefore A is a u-situation and B is a p-situation. Without loss of generality, renumber the variables $\mu_1 \dots \mu_k$ so that $\mu_1 \dots \mu_m$ are situational variables and the rest are not situational variables.

In both halves of the lemma, in order to show that W' is an image of V' we must exhibit an automorphism τ' that establishes this correspondence.

Let us write $\mathrm{PT}(S) = \mathrm{PHYS}(\tau(\mathrm{S}))$, and $S_i = W(\mu_i)$ for $i = 1 \dots m$.

Part 1. There are three cases:

Case A. $m = 0$. In this case, one can choose B = PHYS(A), and τ' to be the identity automorphism.

Case B. Suppose that $A \leqslant S_i$ for some i. Let $\tau' = \tau$, and let B = PT(A). For any j, let S be the LCA of S_j and A. There are four cases:

- B.i. $S_j \leqslant A$. In this case $S = S_j$. Since W is an image of V under τ, $\mathrm{PT}(S) = \mathrm{PT}(S_j) = \mathrm{V}(\mu_j)$.
- B.ii. $A \leqslant S_j$. In this case $S = A$. Since τ is an automorphism, $\tau(S) \leqslant \tau(S_j)$. By Lemma A.2, $\mathrm{PT}(S) = \mathrm{PT}(A) \leqslant \mathrm{PT}(S_j)$ so PT(S) is the LCA of PT(A) and PT(S_j)
- B.iii. A and S_j are unordered but have LCA S. Then S is the LCA of S_i and S_j, so PT(S) is the LCA of PT(S_i) and PT(S_j). Since $\mathrm{PT}(S) < \mathrm{PT}(A) \leqslant \mathrm{PT}(S_i)$, it follows that PT($S$) is the LCA of PT(A) and PT(S_j).
- B.iv. A and S_j have no common ancestor. Hence S_i and S_j have no common ancestor. Hence PT(S_i) and PT(S_j) have no common ancestor. Hence $\mathrm{PT}(A) < \mathrm{PT}(S_i)$ and PT(S_j) have no common ancestor.

Case C. Suppose that A does not precede any of the S_j. Consider the set LL = {LCA(A, S_1) ... LCA(A, S_m)}. If LL is non-empty, let S be the latest situation in LL. We have three cases:

- C.i. LL is empty; that is, none of the S_j are ordered with respect to A. Then none of the values of $\tau(S_j)$ are ordered with respect to τ(A), so by Lemma A.4, none of the values of PT(S_j) are ordered with respect to PT(A). Hence, we may choose $\tau' = \tau$ and B = PT(A).
- C.ii. S is equal to one of the S_i. Then for each S_j, $\mathrm{LCA}(A, S_j) = \mathrm{LCA}(S_i, S_j)$. Thus, again, we may choose $\tau' = \tau$ and $B = \mathrm{PHYS}(\tau(A))$.
- C.iii. S is not equal to any of the S_i. Note that at least there must be one of the $S_j > S$; call this S_x. Let Q be the successor of S such that $Q \leqslant A$. There are two cases:
  - C.iii.a. Q is not a communicative successor of S. Then $\tau(Q)$ is not a communicative successor of $\tau(S)$. For any S_j, if $S < S_j$, let Q_j be the successor of S such that $Q_j \leqslant S_j$. By the construction in Definitions A.9–A.12, it follows that PT(Q) is not equal to PT(Q_j). Therefore PT(S) is the LCA of PT(A) and PT(S_j). If S_j is not ordered with respect to S, then the LCA of S_j and A is the same as the LCA of S_j and S_x (or neither of these LCA's exists), so again $\mathrm{LCA}(\mathrm{PT}(A), \mathrm{PT}(S_j)) = \mathrm{LCA}(\mathrm{PT}(S_x), \mathrm{PT}(S_j)) = \mathrm{PHYS}(\mathrm{LCA}(\tau(S_x), \tau(S_j))) = \mathrm{PHYS}(\mathrm{LCA}(\tau(A), \tau(S_j)))$. Therefore we can choose $\tau' = \tau$ and $B = \mathrm{PHYS}(\tau(A))$.
  - C.iii.b. Q is a communicative successor of S. Here, finally, is the case where τ may need to be modified. Let $Q_1 \ldots Q_p$ be all the successors of S that precede one of the S_i. By property (4) of Definition A.6, there are infinitely many successors of PT(S) that are physically indistinguishable from PT(Q). Let C be one such that is not equal to PT(Q_i) for any i. Let ω be the automorphism of $\mathcal{M}$ that

interchanges the subtree of p-situations following C with the subtree of p-situations following PT(Q) and leaves the rest of $\mathcal{M}$ the same (see Definition A.5). Let $\tau' = \tau \circ \omega$. Let $\mathrm{B} = \mathrm{PHYS}(\tau'(\mathrm{A}))$. Now, suppose $S_j > S$. Then the LCA of S_j and $\mathrm{A} = S$. Since PHYS(τ'(A)) is a descendant of C, which is a successor of PHYS($\tau(S)$) and PHYS($\tau'(S_j)$) is a descendant of PHYS($\tau(Q_j)$ which is a different successor of PHYS($\tau(S)$), it follows that the $\mathrm{LCA}(\mathrm{PHYS}(\tau'(\mathrm{A})), \mathrm{PHYS}(\tau'(S_j))) = \mathrm{PHYS}(\tau(S))$. Alternatively, if S_j is not ordered with respect to S, then we still have $\mathrm{LCA}(\mathrm{PHYS}(\tau(\mathrm{A})), \mathrm{PHYS}(\tau(S_j))) = \mathrm{PHYS}(\mathrm{LCA}(\tau(\mathrm{A}), \tau(S_j)))$, by exactly the same argument as in case C.iii.a.

Part 2. The proof of Part 2 is exactly analogous to that of Part 1, but going in the opposite direction. □

Lemma A.12. *Let α be a prenex formula in $\mathcal{L}$ with m quantifiers and k free variables $\mu_1 \ldots \mu_k$. Let* W *be a valuation from variables $\mu_1 \ldots \mu_k$ to $\mathcal{U}$ and let* V *be an image of* W. *Then α is true under $\mathcal{J}$,* W *if and only if it is true under $\mathcal{I}$,* V.

Proof. By induction on m, the number of quantifiers.

If $m = 0$, then the statement is just Lemma A.10.

Suppose the statement is true for all formulas with m quantifiers. Let α be a formula with $m + 1$ quantifiers. There are four cases:

Case 1: α is true under $\mathcal{J}$, W and α has the form "$\exists_X \beta(X)$", where β is a formula with m quantifiers and $k + 1$ free variables. Since α is true, there exists an entity $A \in \mathcal{U}$ and a valuation $\mathrm{W}' = \mathrm{W} \cup \{X \to A\}$ such that β is true under $\mathcal{J}$, W′. By Lemma A.11 there exists a valuation V′ that is an image of W′. By the inductive hypothesis, β is true under $\mathcal{I}$, V′. Hence α (that is, $\exists_X \beta$) is true under $\mathcal{I}$, V.

Case 2: α is true under $\mathcal{I}$, V and α has the form "$\exists_X \beta(X)$". Since α is true, there exists an entity $B \in \mathcal{M}$ and a valuation $\mathrm{V}' = \mathrm{V} \cup \{X \to B\}$ such that β is true under $\mathcal{I}$, V′. By Lemma A.11 there exists a valuation W′ such that V′ is an image of W′. By the inductive hypothesis, β is true under $\mathcal{J}$, W′. Hence α is true under $\mathcal{J}$, W.

Case 3: α is true under $\mathcal{J}$, W and α has the form "$\forall_X \beta(X)$". Let γ be the transformation into prenex form of $\neg\alpha$. Then γ is false under $\mathcal{J}$, W, and γ has the form "$\exists_X \delta$" where δ is the prenex form of $\neg\beta$. By the contrapositive to case 2 above, γ is false under $\mathcal{I}$, V; hence α is true under $\mathcal{I}$, V.

Case 4: α is true under $\mathcal{I}$, V and α has the form "$\forall_X \beta(X)$". Exactly analogous to case (4), but using the contrapositive to case 1. □

Corollary A.13. *All the physical axioms of $\mathcal{T}$, axioms* T.1–T.7, *and axioms* T.8 *and* T.9 *restricted to physical actions are true in $\mathcal{U}$ under interpretation $\mathcal{J}$.*

Proof. Immediate from Lemma A.12, taking $k = 0$ and using the fact that the axioms in $\mathcal{T}$ and axioms T.1–T.9 are true in $\mathcal{M}$ (by definition of $\mathcal{M}$). □

Lemma A.14. *If* PS1 = PHYS(US1) *and* PS1 *and* PSZ *are ordered, then there exists* USZ *such that* US1 *and* USZ *are ordered, and* PSZ = PHYS(USZ).

Proof. If PS1 = PSZ then USZ = US1.

If PSZ < PS1, then let USZ be the ancestor of US1 at time TIME(PSZ).

If PS1 < PSZ, then let $s_1 = \text{PS1}, s_2 \ldots s_k = \text{PSZ}$ be p-situations such that s_{i+1} is a successor of s_i. Using Definition A.9 iteratively, let US_2 be the simple successor to US1 parallel to PS2, let US_3 be the simple successor to US_2 parallel to PS3, and so on. Then US_k satisfies the desired conditions on USZ. □

Lemma A.15. *Axioms* T.8 *extended to general actions and* K.1–K.8 *are true in* $\mathcal{U}$ *under* $\mathcal{J}$.

(I'm just bunching together the axioms whose proof is easy.)

Proof.

T.8. Immediate from the definition of $\mathcal{J}$("occurs") (Definition A.20).

K.1–K.3. Immediate from Definition A.15, which requires K-ACC(A) to be an equivalence relation on u-situations.

K.4–K.6. Immediate from Definition A.14, which restricts the "possibly accessible" on situations that hold on the left-hand side of each of these relations to those that satisfy the conditions on the right-hand side of these implications; plus Definition A.15, which states that the actual knowledge accessibility relation are a subset of the possibly accessible relations.

K.7,K.8. Immediate from the definition of $\mathcal{J}$("sk_acc") in Definition A.20. □

Lemma A.16. *Axiom* I.1 *is true in* $\mathcal{U}$ *under* $\mathcal{J}$.

Proof. By the definition of $\mathcal{J}$("occurs") in Definition A.20, if occurs(do(AS, inform(AH, Q)),US1, US2) then there exist QA, PF1, USS1, PFA, USSA) such that Q1 = ⟨PF1, USS1⟩, QA = ⟨PFA, USS2⟩, USSA = {US ∈ USS1 | ⟨AS, AH, US1, US⟩ ∈ **k_acc**,}, and ⟨AS, AH, USS2, S1, PHYS(US2)⟩ ∈ MM(US2). Let USQ be the successor of US1 such that USQ ⩽ US2. By Definition A.9, ⟨AS, AH, USS2, S1, PHYS(US2)⟩ ∈ MM(USQ). By Definitions A.10 and A.11, OCCURS(DO(AS, COMMUNICATE(AH)), PHYS(US1), PHYS(US2)). By Definition A.20, occurs(do(AS, communicate(AH)), US1, US2). □

Lemma A.17. *Axiom* T.9 *extended to general actions is true in* $\mathcal{U}$ *under* $\mathcal{J}$.

Proof. Let US1, US2, USX, and USY be u-situations and E an event such that occurs(E, US1, US2), US1 < USX < US2 and USX < USY. Let S1 = PHYS(US1) and S2 = PHYS(US2). By Definition A.20, E is either a physical action or an informative action. The case where E is a physical action is covered in Corollary A.13. Suppose that E is an informative action; let E = ⟨**DO**, AS, ⟨**INFORM**, AH, Q1⟩⟩. By Definition A.20 there exist QA, PF1, USS1, PFA, USSA such that Q1 = ⟨PF1, USS1⟩, QA = ⟨PFA, USS2⟩, USSA = {US ∈ USS1 | ⟨AS, AH, US1, US⟩ ∈ **k_acc**}, and ⟨AS,AH,USSA,S1⟩ ∈ MM(US2). By

Definition A.9, ⟨AS, AH, USSA, S1⟩ ∈ MM(USX). By Axiom T.9 applied to the action DO(AS, COMMUNICATE(AH)) there exists a situation SZ such that ordered(SZ, PHYS(SY)), SZ > SX, and OCCURS(DO(AS, COMMUNICATE(AH), S1, SZ). By Lemma A.14, there exists USZ such that PHYS(USZ) = SZ and USZ is ordered with respect to USY. It follows that USZ > USX and that ⟨AS, AH, USS2, S1⟩ ∈ MM(USZ). By Definition A.20, occurs(do(AS, inform(AH, Q)), US1, USZ). □

Lemma A.18. *Axiom* I.2 *is true in* $\mathcal{U}$ *under* $\mathcal{J}$.

Proof. Let AS, AH be agents, let US1, US2 be u-situations, and let Q = ⟨PF, USSQ⟩ be a general fluent. Let US1ACC = {USA | ⟨AS, AH, US1, USA⟩ ∈ **sk_acc**}, the set of situations accessible from US1 in the shared knowledge of AS and AH. Let USSA = USSQ∩US1ACC, the set of situations satisfying Q that are knowledge accessible from S1, relative to the shared knowledge of AS and AH. Let S1 = PHYS(US1).

Suppose that occurs(do(AS, inform(AH, Q)), US1, US2). By Definition A.20 (denotation of "occurs"), the tuple ⟨AS, AH, USSA, S1⟩ ∈ MM(US2). Let USY be the successor of US1 that is an ancestor of US2. By Definitions A.9, A.11, and A.12 it follows that MM(USY) contains the tuple ⟨AS, AH, USSA, S1⟩. By Definition A.10, USSA is a possible communicative content for S1 from AS to AH; hence, by Definition A.10, every situation that is knowledge accessible from US1 relative to AS is an element of USSA and therefore an element of USSQ ⊃ USSA. By Definition A.20 ("holds") Q holds in every situation accessible from US1.

Conversely, if Q holds in every situation accessible from S1, then USSA is a possible communicative content from AS to AH. Suppose that OCCURS(DO(AS, COMMUNICATE(AH)), S1, S2). Let SY be the successor of S1 such that SY ⩽ S2. By Definition A.12, there exists an informative successor USY of US1 such that ⟨AS, AH, USSA, S1⟩ ∈ MM(USY). By Axiom T.9 there exists a situation USZ ⩾ USY such that OCCURS(DO(AS, COMMUNICATE(AH)), US1, USZ). By Definitions A.9, A.11, A.12 ⟨AS, AH, USSA, S1⟩ ∈ MM(USZ). By Definition A.20, occurs(do(AS, inform(AH,Q)), S1, SZ). □

Lemma A.19. *Axiom* I.3 *is true in* $\mathcal{U}$ *under* $\mathcal{J}$.

Proof. Assume that occurs(do(AS, inform(AH, Q)), US1, US2) and that k_acc(AH, US2, US2A). We need to prove that there exists a situation US1A such that occurs(do(AS, inform(AH, Q)), US1A, US2A) and k_acc(AH, US1, US1A).

Define USSA as in the proof of Lemma A.18. By Definition A.20 (denotation of "occurs") since occurs(do(AS, inform(AH, Q)), US1, US2) it follows that the tuple ⟨AS, AH, USSA, PHYS(US1)⟩ ∈ MM(US2) and OCCURS(DO(AS, COMMUNICATE(AH)), PHYS(US1), PHYS(US2)). By Definition A.15, since k_acc(AH, US2, US2A), US2A is possibly knowledge accessible from US2 relative to AH. By Definition A.14, the tuple ⟨AS, AH, USSA, PS1A⟩ ∈ MM(US2A) for some p-situation PS1A < PHYS(US2A), and OCCURS(DO(AS, COMMUNICATE(AH)), PS1A, PHYS(US2A)). By Theorem 3 and axiom K.8, any two situations that are sk_acc are at the same time. Hence, all the situations in USSA are at the same time, and by Definition A.10

this time must be equal to TIME(US1) and to TIME(US1A). Hence TIME(US1) = TIME(US1A). By axiom A.4, since k_acc(AH, US2, US2A), US1 < US2, US1A < US2A and TIME(US1) = TIME(US1A), it follows that k_acc(AH, US1, US1A). Hence, the set of situations that are accessible relative to the shared knowledge of AS and AH is the same starting from US1 as starting from US1A. Hence the act of AS informing AH of Q starting in US1A uses the tuple ⟨AS, AH, USSA, PS1A⟩. Thus by Definition A.20, occurs(do(AS, inform(AH, Q)), US1A, US2A). □

Lemma A.20. *Axiom* I.4 *is true in* $\mathcal{U}$ *under* $\mathcal{J}$.

Proof. Suppose that occurs(do(AS, inform(AH, QX)), US1, US2) and that QY is a fluent. Let US3 be the successor of US1 such that US3 ⩽ US2. Let

$$\mathrm{QX} = \langle \mathrm{PFX}, \mathrm{USSQX} \rangle$$

$$\mathrm{QY} = \langle \mathrm{PFY}, \mathrm{USSQY} \rangle$$

$$\mathrm{QXA} = \mathrm{USSQX} \cap \{\mathrm{USA} \mid \mathrm{sk\_acc}(\mathrm{AS}, \mathrm{AH}, \mathrm{US1}, \mathrm{USA})\} \quad \text{and}$$

$$\mathrm{QYA} = \mathrm{USSQY} \cap \{\mathrm{USA} \mid \mathrm{sk\_acc}(\mathrm{AS}, \mathrm{AH}, \mathrm{US1}, \mathrm{USA})\}$$

By Definition A.20, ⟨AS, AH, QXA, PHYS(US1)⟩ ∈ MM(US2). By Definitions A.9, A.11, A.12, ⟨AS, AH, QXA, PHYS(US1)⟩ ∈ MM(US3).

I. (Left to right in the two-way implication.)
Suppose that occurs(do(AS, inform(AH, QY)), US1, US2). By the same argument as above ⟨AS, AH, QYA, PHYS(US1)⟩ ∈ MM(US3). But by Definition A.11, US3 contains at most one inform indicator with starting point PHYS(US1), speaker AS, and hearer AH. Hence QXA = QYA. That is, if situation USA is accessible from US1 relative to the shared knowledge of AS and AH, then QX holds in USA iff QY holds in USA.

II. (Right to left in the two-way implication.)
If it is the case that

$$\forall_{S1A}\ \mathrm{sk\_acc}(\mathrm{AS}, \mathrm{AH}, \mathrm{S1}, \mathrm{S1A}) \Rightarrow [\mathrm{holds}(S1A, \mathrm{QX}) \Leftrightarrow \mathrm{holds}(S1A, \mathrm{QY})]$$

then QXA = QYA, so by Definition A.20, occurs(do(AS, inform(AH, QY)), US1, US2). □

Lemma A.21. *Axiom* I.5 (*the comprehension axiom*) *is true in* $\mathcal{U}$ *under* $\mathcal{J}$.

Proof. Immediate from Definitions A.17 and A.20, using the comprehension axiom of Zermelo–Fraenkel set theory (also known as the “subset” or “separation” axiom). Since there exists a well-defined set U-SITS of all situations, the ZF axiom asserts that every such formula defines a subset of U-SITS. See, for example, [15, p. 36]. □

Considering how problematic the comprehension axiom would seem to be it may be surprising that it has a one-step proof. In fact, one might say that the whole construction we went through in Section A.3 is precisely tailored so that the comprehension axioms *should* have a one-step proof. Nonetheless the reader may well have legitimate worries about such a powerful axiom, that are hardly assuaged by the above proof. Let me therefore discuss further how this whole construction works.

The key point is this: There is *no circularity whatever* in the whole structure of definitions given in Section 3. The structure of u-situations is built up iteratively forward in time. The label on an "inform" action A is a set of u-situations contemporaneous with the start of A; it gives rise to a new u-situations at the next point in time. Iterating from 1 to infinity gives us a well-defined and fixed set $\mathcal{U}$ of all u-situations. Definition A.17 defines a fluent as a subset of $\mathcal{U}$. Definition A.20 defines the occurrence of an inform action in terms of these fluents and of the labels on the actions. More generally, Definition A.20 defines the denotation of every symbol in $\mathcal{W}$ extensionally, in terms of structures over $\mathcal{U}$ and $\mathcal{M}$ and the interpretation $\mathcal{I}$; no aspect of $\mathcal{J}$ is defined in terms of $\mathcal{J}$ itself (except as a convenient abbreviation). Having adopted Definition A.20, $\mathcal{J}$ is now fixed, and it is fixed which fluents satisfy which formulas under $\mathcal{J}$.

But is not it inherently circular to say, for example,

q1 is the fluent such that

$$\forall_S \text{ holds}(S, \text{q1}) \Leftrightarrow \exists_{AS,AH,S2,Q} \text{ occurs}(\text{do}(AS, \text{inform}(AH, Q)), S, S2)$$

considering that the quantification over Q contains q1 itself? Not at all, no more than saying

0 is the number such that, $\forall_X X + 0 = X$

when the quantification over X includes 0 itself. The formula above is just a description of q1, and the axioms are sufficient to guarantee that a q1 satisfying this definition exists.

Theorem 1. *Let $\mathcal{T}$ be an acceptable physical theory, and let $\mathcal{A}$ be $\mathcal{T}$ together with axioms* K.1–K.8 *and* I.1–I.5*, and with* T.8 *and* T.9 *extended to arbitrary actions. Then $\mathcal{A}$ is consistent.*

Proof. We have shown that a model and an interpretation satisfying $\mathcal{A}$ can be constructed. □

Theorem 2. *Let $\mathcal{T}$ be an acceptable physical theory, and let $\mathcal{U}$ be the union of*:

A. $\mathcal{T}$.
B. *Axioms* K.1–K.7 *and* I.1–I.5.
C. *A collection of domain-specific knowledge acquisition axioms of the form specified in Section* 3.5.
D. *The frame axiom* I.6 *associated with the axioms in* (C).
E. *Any set of axioms $\mathcal{K}$ specifying the presence or absence of k_acc relations among situations at time* 0 *as long as*:
   i. *The axioms in $\mathcal{K}$ do not refer to any situations of time later than* 0.
   ii. *The axioms in $\mathcal{K}$ are consistent with $\mathcal{T}$, axioms* K.1–K.3, K.5 (*as regards knowing the feasibility of actions at time* 0), *and the axioms in* (C).

Then $\mathcal{U}$ is consistent.

Proof (*Sketch*). The proof of Theorem 1 needs to be modified as follows:

- In Definition A.8, initialize the K_ACC function at time 0 to satisfy the union of the axioms in (E) with the axioms enumerated in E.ii.
- In Definition A.14, add to the conditions on US1B being possibly knowledge accessible from US1A:

  For each axiom in (C) of the form "A always knows whether $\Phi_i(A, S)$," the condition $\Phi_i(US1B) \Leftrightarrow \Phi_i(US1A)$ must hold.

- Modify the second bullet in Definition A.15 to read, "For each agent A, K_ACC1(A) is the relation over u-situations, 'US1B is knowledge accessible from US1A relative to A' ".

The proof that the additional axioms enumerated in Theorem 2 are satisfied is then straightforward. □

References

[1] A. Baltag, L. Moss, S. Solecki, The logic of public announcements: Common knowledge and private suspicions, Unpublished, 2002, http://www.cs.indiana.edu/ftp/techreports/TR534.html.

[2] J. van Benthem, 'One is a Lonely Number': On the logic of communication, ILLC Tech Report 2003-07, Institute for Logic, Language and Computation, University of Amsterdam, 2003.

[3] D. Chapman, Planning for conjunctive goals, Artificial Intelligence 32 (1987) 333–378.

[4] P.R. Cohen, C.R. Perrault, Elements of a plan-based theory of speech acts, Cognitive Sci. 3 (3) (1979) 177–212.

[5] P.R. Cohen, H. Levesque, Intention is choice with commitment, Artificial Intelligence 42 (2–3) (1990) 213–261.

[6] M. Colombetti, A modal logic of intentional communication, Math. Social Sci. 38 (1999) 171–196.

[7] E. Davis, Inferring ignorance from the locality of visual perception, in: Proceedings of AAAI-88, St. Paul, MN, AAAI Press/MIT Press, Cambridge, MA, 1988, pp. 786–790.

[8] E. Davis, Representations of Commonsense Knowledge, Morgan Kaufmann, San Mateo, CA, 1990.

[9] E. Davis, A first-order theory of communicating first-order formulas, in: D. Dubois, C. Welty, M. Williams (Eds.), Proceedings of the Ninth International Conference on Principles of Knowledge Representation and Reasoning (KR2004), Whistler, BC, AAAI Press, Menlo Park, CA, 2004, pp. 235–245.

[10] E. Davis, L. Morgenstern, A first-order theory of communication and multi-agent plans, J. Logic Comput., in press.

[11] R. Fagin, J. Halpern, M. Vardi, What can machines know? On the properties of knowledge in distributed systems, J. ACM 39 (1992) 328–376.

[12] R. Fagin, J. Halpern, Y. Moses, M. Vardi, Reasoning about Knowledge, MIT Press, Cambridge, MA, 1995.

[13] T. Finin, J. Weber, M. Genesereth, D. McKay, J. McGuire, R. Pelavin, S. Shapiro, C. Beck, Specification of the KQML agent communication language, DARPA Knowledge Sharing Initiative External Interfaces Working Group, 1993.

[14] FIPA, The foundation for intelligent physical agents, http://www.fipa.org/, 2001.

[15] A. Fraenkel, Y. Bar-Hillel, A. Levey, A Foundations of Set Theory, second ed., Noord-Hollandsche U.M., Amsterdam, 1973.

[16] M. Gardner, The Unexpected Hanging and Other Mathematical Diversions, Chicago University Press, Chicago, IL, 1991.

[17] J. Halpern, A theory of knowledge and ignorance for many agents, J. Logic Comput. 7 (1) (1997) 79–108.

[18] J. Halpern, M. Vardi, The complexity of reasoning about knowledge and time in asynchronous systems, in: Proc. 20th ACM Symp. on Theory of Computation, 1988, pp. 53–65.
[19] J. Halpern, M. Vardi, The complexity of reasoning about knowledge time, I: lower bounds, J. Comput. Systems Sci. 38 (1) (1989) 195–237.
[20] J. Halpern, M. Vardi, Model checking vs. theorem proving: A manifesto, in: J. Allen, R. Fikes, E. Sandewall (Eds.), Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR1991), Morgan Kaufmann, San Mateo, CA, 1991, pp. 325–334.
[21] P. Hayes, The naive physics manifesto, in: D. Michie (Ed.), Expert Systems in the Micro-Electronic Age, Edinburgh University Press, Edinburgh, 1978.
[22] H. Levesque, All I know: A study in auto-epistemic logic, Artificial Intelligence 42 (1990) 263–309.
[23] A. Lomuscio, M. Ryan, A spectrum of modes of knowledge sharing between agents, in: N. Jennings, Y. Lespérance (Eds.), Intelligent Agents VI: Agent Theories, Architectures, and Languages, in: Lecture Notes in Artificial Intelligence, vol. 1757, Springer, Berlin, 2000, pp. 13–26.
[24] J. McCarthy, Programs with common sense, in: M. Minsky (Ed.), Semantic Information Processing, MIT Press, Cambridge, MA, 1968, pp. 403–418.
[25] J. McCarthy, P. Hayes, Some philosophical problems from the standpoint of artificial intelligence, in: B. Meltzer, D. Michie (Eds.), Machine Intelligence, vol. 4, Edinburgh University Press, Edinburgh, 1969, pp. 463–502.
[26] D. McDermott, A temporal logic for reasoning about processes and plans, Cognitive Sci. 6 (1982) 101–155.
[27] R. van der Meyden, K. Wong, Complete axiomatizations for reasoning about knowledge and branching time, Studia Logica 75 (1) (2003) 93–123.
[28] R. Moore, Reasoning about Knowledge and Action, Tech. Note, vol. 191, SRI International, Menlo Park, CA, 1980.
[29] R. Moore, A formal theory of knowledge and action, in: J. Hobbs, R. Moore (Eds.), Formal Theories of the Commonsense World, ABLEX Publishing, Norwood, NJ, 1985, pp. 319–358.
[30] R. Moore, Semantical considerations on nonmonotonic logic, Artificial Intelligence 25 (1985) 75–94.
[31] L. Morgenstern, Foundations of a logic of knowledge, action, and communication, PhD thesis, NYU, 1988.
[32] R. Parikh, R. Ramanujam, A knowledge-based semantics of messages, J. Logic Language Inform. 12 (4) (2003) 453–467.
[33] J. Plaza, Logics of public announcements, in: Z.W. Ras (Ed.), Proc. 4th International Symposium on Methodologies for Intelligent Systems, Charlotte, NC, North-Holland, Amsterdam, 1989.
[34] W.V.O. Quine, On a so-called paradox, Mind 62 (1953) 65–67.
[35] A.S. Rao, Decision procedures for propositional linear time belief-desire-intention logics, in: M. Wooldridge, J. Müller, M. Tambe (Eds.), Intelligent Agents II: Agent Theories, Architectures, and Languages, in: Lecture Notes in Artificial Intelligence, vol. 1037, Springer, Berlin, 1995, pp. 33–48.
[36] R. Reiter, Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems, MIT Press, Cambridge, MA, 2001.
[37] M.D. Sadek, P. Bretier, F. Panaget, ARTIMIS: Natural dialogue meets rational agency, in: M. Pollack (Ed.), Proceedings of IJCAI-97, Nagoya, Japan, Morgan Kaufmann, San Francisco, CA, 1997, pp. 1030–1035.
[38] R. Scherl, H. Levesque, The frame problem and knowledge producing actions, in: Proceedings of AAAI-93, AAAI Press/MIT Press, Cambridge, MA, 1993, pp. 689–695.
[39] R. Scherl, H. Levesque, Knowledge, action, and the frame problem, Artificial Intelligence 144 (1) (2003) 1–39.
[40] M. Wooldridge, An Introduction to MultiAgent Systems, Wiley, New York, 2002.
[41] M. Wooldridge, A. Lomuscio, Reasoning about visibility, perception, and knowledge, in: N. Jennings, Y. Lespérance (Eds.), Intelligent Agents VI: Agent Theories, Architectures, and Languages, in: Lecture Notes in Artificial Intelligence, vol. 1757, Springer, Berlin, 2000, pp. 1–12.