Mutual Belief Revision (Preliminary Report)

Ron van der Meyden
NTT Basic Research Labs
3-1 Wakamiya Morinosato Atsugi-shi
Kanagawa 243-01, Japan

Abstract

A semantical theory of belief revision in synchronous multi-agent systems is presented. The theory assumes that the world is static, and that the belief revision methods used by the agents are common knowledge. It is shown that the theory has as a special case the revision of knowledge in a broadcast variant of Fagin and Vardi's *communicating scientists* model, in which knowledge is defined in terms of an equivalence relation on runs in a distributed system. Next, the theory is applied to give a semantic reconstruction of a default logic theory of speech acts proposed by Perrault. This reconstruction characterizes the way the speech act transforms an *only knowing* representation of the agent's belief states. The results are shown to differ from those obtained by Perrault. In particular, it is argued that a sincere assertion does not necessarily lead to mutual belief of the proposition asserted, even when the hearer does not initially have any beliefs about this proposition.

1 INTRODUCTION

The question of how an agent should maintain its epistemic state has long been a subject of interest in the areas of artificial intelligence, databases and philosophy, and the literature on this topic is extensive. In particular, there has been a considerable amount of work on the semantics of single agent belief revision [Gärdenfors, 1986; Katsuno and Mendelzon, 1991]. In multiple agent situations, belief revision is more complex, since an agent must revise not just its own beliefs about the world, but also its beliefs about other agents' beliefs about the world. Moreover, an agent must also revise its beliefs about other agents' beliefs about its own beliefs, and so on. We may call this iterated process *mutual belief revision*. The extant theories treating the dynamics of belief in multiple agent environments are primarily syntactical [Appelt and Konolige, 1988; Cohen and Levesque, 1990; Isozaki and Shoham, 1992; Morgenstern, 1990; Perrault, 1990; Thomason, 1990]. In this paper we develop a semantic theory of mutual belief revision.

Among the domains of application of such a theory is a strand in natural language pragmatics that seeks to characterize the meaning of a speech act by describing the effect it has on the mental states of the participants of the conversation [Allen and Perrault, 1980; Cohen and Levesque, 1990; Cohen and Perrault, 1979; Perrault, 1990]. Another potential application is the semantics of agent/knowledge oriented programming languages [Fagin *et al.*, 1994; Moses and Kislev, 1993; Shoham, 1993], which include as primitive constructs modal operators for knowledge and belief.

One can study the dynamics of belief under a variety of assumptions about the nature of agents and their environment. We work in this paper with a very specific set of assumptions. Prototypical of the situations to which our theory is intended to apply are face to face communication, mutual observation of a set of circumstances, and broadcast in synchronous systems with guaranteed message delivery. Thus, we assume agents revise their beliefs synchronously, in response to an event whose occurrence is common knowledge. Each agent is assumed to employ an individual belief revision method, which is common knowledge to the other agents.

In addition, we assume that the world is static - it is only the beliefs of the agents that change. In the terminology of [Katsuno and Mendelzon, 1991], we deal with the *revision* of belief, occasioned when additional information is obtained, or some existing beliefs are are found to be false. This is opposed to the notion of *update*, applicable when a change of belief is required because of a change in the state of the world.[1]

[1] For an investigation of update in multi-agent environments see [Meyden, 1994].

The theory we develop represents mutual belief revision as an operator on certain sets of Kripke structures that are canonical for the multi-agent version of the perfect introspection logic K45. However, because our theory is semantical it applies also to logics stronger than K45, such as KD45 and S5. The mutual belief revision process is factored into two stages. In the first stage, agents revise their individual beliefs, ignoring the fact that other agents are also revising their beliefs. In the second stage, the other agents are taken into account. The main concern of the paper is with the relationship between the individual belief revision methods and the agents' mutual belief revision.

The model of individual belief revision used in this paper is very general, and does not satisfy the postulates for rational belief revision recommended by [Alchourrón *et al.*, 1985], commonly referred to as the *AGM postulates*. This is in part because our focus is on the issue of mutuality rather than rationality, and we prefer to cast the theory at the most general possible level. However, we would also argue that the AGM postulates, and others like them, are too restrictive. Indeed, we present an example of interest in mutual belief revision (derived in essence from work of Perrault) in which the AGM postulates are not satisfied by the revision method of the individual agents. Nevertheless, one could restrict our theory by adding the AGM postulates as constraints on the agents' individual belief revision methods. One of the merits of our theory is that it allows for the incorporation of such assumptions in a very natural way.

In Section 3 capture the above assumptions on agents and their environment by formalizing the relationship between mutual belief revision and individual belief revision. We show that for each assignment of agents' individual belief revision methods there is a unique mutual belief revision operator satisfying this formalization. The remainder of the paper is devoted to justifying this theory by studying its consequences in two special cases.

First, we consider knowledge, a special case of belief. The dynamics of knowledge are better understood than those of belief: there exists an independently motivated account of the revision of knowledge, arising directly from the semantics of knowledge in distributed systems [Halpern and Moses, 1990]. We consider a particular distributed system, the *scientists in conference* model, which satisfies the assumptions above. This system is a modification of the *communicating scientists* model of [Fagin *et al.*, 1994; Fagin and Vardi, 1986], which was based on message passing. We show that when our theory of mutual belief revision is applied to the states of knowledge arising in the scientists in conference system, the results are identical to those implied by the semantics of knowledge. We take this to be strong evidence for the reasonableness of our approach.

As a second application, we consider the implications of the theory for an account of speech acts proposed by Perrault. Perrault's work was syntactical, being cast in terms of an extension of default logic [Reiter, 1980]. In order to give a semantic reconstruction, we use the notion of *only knowing* [Levesque, 1990] to represent the agents' initial state of knowledge. We then study the prediction of our theory for the state of knowledge after an assertion. The results turn out to differ from those obtained by Perrault. In particular, our theory entails that a sincere assertion does not necessarily lead to mutual belief of the proposition asserted, even when the hearer does not initially have any beliefs about this proposition. However, a weaker proposition does become mutual knowledge, and we argue that this is an appropriate result, that again supports the reasonableness of our theory.

The structure of the paper is as follows. The preliminary Section 2 introduces a particular class of canonical Kripke structures for $K45_n$. Section 3 describes our theory of mutual belief revision. Section 4 shows that the theory correctly describes the revision of knowledge in the scientists in conference model. In Section 5 we apply the theory to model Perrault's default logic theory of speech acts. Section 6 concludes.

2 PRELIMINARIES

We work with a propositional modal language $\mathcal{L}_n$ generated over the set of propositional constants Φ. The language contains a modal operator B_i for each agent $i = 1 \ldots n$, representing the belief of agent i. The set of formulae is the smallest set containing Φ such that if φ and ψ are formulae then $\neg\varphi$, $\varphi \wedge \psi$ and $B_i\varphi$ are formulae. A formula is said to be *i-objective* if it is a boolean combination of atomic formulae and formulae of the form $B_j\varphi$ where $j \neq i$.

The class of semantic structures used to interpret this language will be the $K45_n$ structures, defined as follows. A *propositional assignment* is a function $\alpha : \Phi \rightarrow \{0, 1\}$. A *valuation* on a set X is a function V mapping each element of X to a propositional assignment. A $K45_n$ structure is a tuple $M = \langle W, R_1, \ldots, R_n, V\rangle$ where W is a set of worlds, the R_i are binary relations on W, and V is a valuation on W. The relations R_i are required to be *transitive* and *euclidean* [Chellas, 1980]. An alternate way to understand these conditions is as follows. Given a world w and agent i, define the *possibilities of i at w* to be the set $poss_i(w) = \{v \in W \mid wR_iv\}$. Then R_i is transitive and euclidean just when for all worlds w and all $v \in poss_i(w)$, we have $poss_i(v) = poss_i(w)$. In particular, note that this means that R_i restricted to $poss_i(w)$ is an equivalence relation.

A $K45_n$ *situation* is a pair (M, w) where M is a $K45_n$ structure and w is a world of M. Satisfaction of formulae of $\mathcal{L}_n$ in situations is denoted by a binary relation symbol '$\models$', and defined along the usual lines:

$(M, w) \models p$ if $V(w)(p) = 1$, where $p \in \Phi$,

$(M, w) \models \varphi \wedge \psi$ if $(M, w) \models \varphi$ and $(M, w) \models \psi$,

$(M, w) \models \neg\varphi$ if not $(M, w) \models \varphi$,

$(M, w) \models B_i\varphi$ if $(M, w') \models \varphi$ for all $w' \in W$ such that wR_iw'.

Intuitively, a situation describes a particular state of the world, the agents' beliefs about the world, their beliefs about each others beliefs, etc. It is well-known that requiring the relations R_i to be transitive and euclidean results in the validity of the formulae $B_i\varphi \rightarrow B_iB_i\varphi$ and $\neg B_i\varphi \rightarrow B_i\neg B_i\varphi$, known respectively as the *positive* and *negative introspection* rules.

We will, in the next section, seek to represent mutual belief revision by means of an operator mapping situations to situations. Since the set of worlds of a Kripke structure may have arbitrarily large cardinality, the class of all situations does not form a set, which means that a function from all situations to situations cannot be formulated within set theory. To get around this problem, and to enable explicit description of the revision operators, we will use certain canonical structures. These will be infinite tree-like structures, in which the degree of branching is bounded.

It is convenient to represent these structures using a generalization of the notion of tree domain. Let L be a set of *labels*, equipped with a valuation V. Define an (L, n)*-sequence of length* m to be a sequence of the form $l_0i_1l_1i_2l_2 \ldots i_ml_m$ where $m \geq 0$, the l_j are labels in L, the i_j are agents in $\{1 \ldots n\}$ and $i_j \neq i_{j+1}$ for each $j = 1 \ldots m-1$. A set T of (L, n)-sequences is said to be *closed under prefixes* if whenever $l_0i_1l_1i_2l_2 \ldots i_nl_m \in T$ we have $l_0i_1l_1i_2l_2 \ldots i_kl_k \in T$ for all $k \leq m$. An (L, n)*-labelled tree* is a possibly infinite set T of (L, n)-sequences that is closed under prefixes and contains a unique (L, n)-sequence of length 0, i.e. a label l_0 in L. In this case we define $root(T) = l_0$. We write $\mathcal{T}(L, n)$ for the set of (L, n)-labelled trees on L. When L and n are clear from the context we will refer simply to "labelled trees".

We call the elements of a labelled tree T the *vertices* of T. If v is a vertex of T and the sequence[2] $w = v \cdot il$ is also a vertex of T, where $i \in \{1 \ldots n\}$ and $l \in L$, then we say that w is an i*-successor* of v. An (L, n)-labelled tree T has an obvious interpretation as a tree with edges labelled by elements of $\{1 \ldots n\}$, vertices labelled by elements of L. We define the labelling function lab from n-sequences to L by $lab(l_0i_1l_1i_2l_2 \ldots i_ml_m) = l_m$. Note that no two distinct i-successors of any vertex have the same label. Thus, the cardinality of branching is bounded by the cardinality of the set L. Additionally, note that the requirement that $i_j \neq i_{j+1}$ in the definition of (L, n)-sequences implies that an i-successor of any vertex can have no i-successors itself, though it may have j-successors for $j \neq i$. The reason for this will become apparent below.

[2]The operator '$\cdot$' denotes concatenation of sequences.

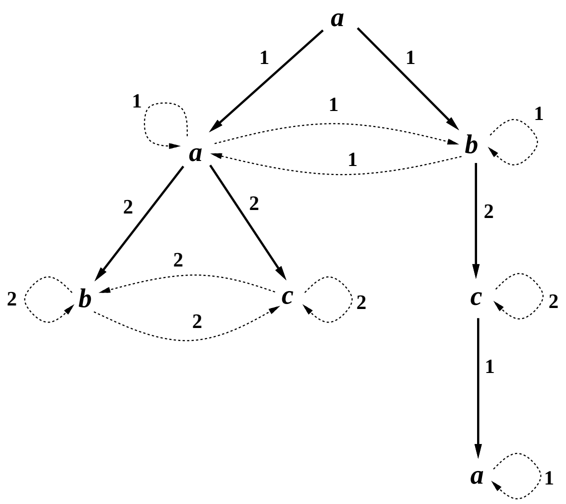

Figure 1: A labelled tree

A labelled tree in which the root has no i-successors will be called an i*-objective tree*. We write $\mathcal{T}_i(L, n)$ for the set of i-objective (L, n)-labelled trees. An (L, n)*-labelled forest* is a set F of (L, n)-labelled trees such that when T and T' are distinct trees in F we have $root(T) \neq root(T')$. We say that the forest F is i-objective if every tree in F is i-objective, and write $\mathcal{F}_i(L, n)$ for the set of i-objective (L, n)-labelled forests.

If T is a labelled tree and v is a vertex of T with $lab(v) = l_0$ then we define the *subtree of* T *rooted at* v, denoted by $T \uparrow v$, to be the tree containing the sequences $l_0i_1l_1i_2l_2 \ldots i_ml_m$ such that T contains the sequence $v \cdot i_1l_1i_2l_2 \ldots i_ml_m$. An i*-child* of a vertex v in a labelled tree T is a labelled tree $T \uparrow w$ where w is an i-successor of v. We write $child_i(T)$ for the set of all i-children of the root of T. Note that $child_i(T)$ is an i-objective forest for all trees T.

Example 2.1: Suppose $L = \{a, b, c\}$. Then

$$T = \{a,\ a1a,\ a1a2b,\ a1a2c,\ a1b,\ a1b2c,\ a1b2c1a\}$$

is an $(L, 2)$-labelled tree, depicted in Figure 1. Here the solid edges represent the tree structure - the meaning of the broken edges will be explained below. If $v = a1a$ then $T \uparrow v = \{a,\ a2b,\ a2c\}$. Note that T is a 2-objective tree. The set of 1-children of T is the 1-objective forest $child_1(T) = \{\{a,\ a2b,\ a2c\},\ \{b,\ b2c,\ b2c1a\}\}$.

An (L, n)-labelled tree T may be regarded as a K45$_n$ structure $M(T) = \langle W, R_1, \ldots, R_n, V_M\rangle$. The worlds

W of this structure are just the vertices of T, and the valuation V_M is given by $V_M(v) = V(lab(v))$, where v is a vertex of T and V is the valuation on L. The accessibility relations R_i are not the i-successor relations, however, but the larger relations given by vR_iv' if and only if v' is an i-successor of v or there exists a vertex $w \in T$ such that v and v' are both i-successors of w. Note that because no i-successor of a vertex has an i-successor, these relations are transitive and euclidean, so $M(T)$ is indeed a K45$_n$ structure.

By adding to $M(T)$ the world represented by root of the tree T, we also obtain a K45$_n$ situation $(M(T), root(T))$. We write $T \models \varphi$ when $(M(T), root(T)) \models \varphi$. Note that under this interpretation of T the i-objective forest $child_i(T)$ constitutes a set of situations representing agent i's state of belief.

Example 2.2: Consider the tree of example 2.1. The broken edges in Figure 1 represent the additional tuples added to the tree edges to construct the relations R_i. If Φ contains a single proposition p and the valuation V on L is given by $V(a)(p) = V(b)(p) = 0$ and $V(c)(p) = 1$, then $T \models B_1 \neg B_1 B_2 p$.

Conversely, the class of all labelled trees is able to represent all K45$_n$ situations by a construction similar to a well known 'unravelling' construction [Bull and Segerberg, 1984]. Suppose we are given a structure $M = \langle W, R_1, \ldots, R_n, V\rangle$ and a world w of M. Let the set L of labels be the set W of worlds of M, equipped with the valuation V of M. Take $T(M,w)$ to be the (L,n)-labelled tree containing a vertex $wi_1u_1i_2u_2\ldots i_mu_m$ for every sequence of worlds $u_1 \ldots u_m$ and every sequence of indices $i_1 \ldots i_m$ without consecutive repetitions such that $wR_{i_1}u_1R_{i_2}u_2\ldots R_{i_m}u_m$.

This construction is unlike the standard one in that it deals with multi-agent structures, but the main difference is the condition that there be no consecutive repetitions of agent indices. Intuitively, the reason for this is that, because we are dealing with K45$_n$ structures, allowing agent repetitions yields trees with a great deal of redundancy, satisfying certain constraints related to the positive and negative introspection conditions. We could formulate our work in this paper in terms of such trees, but find it more convenient to eliminate the need to deal explicitly with the additional constraints by factoring out the redundant parts of the tree. The following result shows that, with respect to the language $\mathcal{L}_n$, the tree $T(M,w)$ expresses the same information as the situation (M,w).

Proposition 2.3: If $\varphi \in \mathcal{L}_n$ then $(M,w) \models \varphi$ iff $T(M,w) \models \varphi$.

This result shows that labelled trees are canonical in that they are able to represent all situations. In some respects the labelled trees form too large a class of structures, however. Suppose the tree T' is obtained from the tree T by replacing the label l at the root by another label l' associated with the same assignment, i.e. for which $V(l) = V(l')$. Since the vertex labelling is opaque to the language $\mathcal{L}_n$, these trees satisfy the same formulae, and intuitively represent the same information about the state of the world and the beliefs of the agents. Further, suppose that T_1 and T_1' are trees identical except in that the 1-children of T_1 are T and T', whereas T_1' has only the 1-child T. Then the trees T_1 and T_1' again satisfy the same formulae: intuitively, the fact that the situation corresponding to T is a possibility for agent 1 is redundantly represented in the tree T_1.

A general notion that captures this sort of invariance is the following. We say that two trees $T, T' \in \mathcal{T}(L,n)$ are *congruent*, and write $T \cong T'$, if there exists a binary relation R on $T \times T'$ such that $root(T) R root(T')$ and for all vertices $u \in T$ and $v \in T'$, and all agents $i = 1 \ldots n$,

1. if uRv then $V(lab(u)) = V(lab(v))$, and
2. if uRv and u' is an i-successor of u in T then there exists an i-successor v' of v in T' such that $u'Rv'$, and
3. if uRv and v' is an i-successor of v in T' then there exists an i-successor u' of u in T such that $u'Rv'$.

If the relation R is additionally a one-to-one correspondence, then we say that the trees T, T' are *isomorphic*. We say that two forests F and F' are congruent if for all trees $T \in F$ there exists a tree $T' \in F'$ with $T \cong T'$, and vice versa.

Example 2.4: Suppose that the valuation V on the set L of Example 2.1 satisfies $V(b) = V(c)$. Then the tree

$$T' = \{a,\ a1a,\ a1a2b,\ a1c,\ a1c2b,\ a1c2b1a\}$$

is congruent to the tree T of Example 2.1 by means of the relation R consisting of the tuples

$$\begin{array}{ll} \langle a,\ a\rangle & \\ \langle a1a,\ a1a\rangle & \langle a1b,\ a1c\rangle \\ \langle a1a2b,\ a1a2b\rangle & \langle a1b2c,\ a1c2b\rangle \\ \langle a1a2c,\ a1a2b\rangle & \langle a1b2c1a,\ a1c2b1a\rangle. \end{array}$$

Conditions 1-3 of the definition of congruence generalize the notion of a *zig-zag connection* [Benthem, 1984] between Kripke structures to the multi-agent case, and are also related to the *bisimulations* [Hennessy and Milner, 1985] used in the theory of concurrency. The first part of the following proposition, which is similar to known results for these related notions, may be interpreted as stating that congruent trees represent the same state of mutual belief.

Proposition 2.5:

1. If $T \cong T'$ then for all formulae $\varphi \in \mathcal{L}_n$ we have $T \models \varphi$ iff $T' \models \varphi$.

2. If $T \cong T'$ then $child_i(T) \cong child_i(T')$ for each agent i.

We will take the point of view in what follows that congruent trees should be revised in a similar fashion.

3 A Theory of Mutual Belief Revision

We now present our theory of mutual belief revision. Our approach will be indirect. Instead of explicitly describing the operator representing the mutual belief revision process, we begin by stating a number of constraints on this operator. We then derive the operator from these constraints, by showing that they have an essentially unique solution.

Prototypical of the sort of situation we wish to model is an utterance made in a face-to-face conversation, where all agents hear the utterance, each observes that the others are in a position to hear the utterance, and each observes the others making this observation. Informally, in such situations agents simultaneously revise their beliefs, and the fact that belief revision is taking place is common knowledge. The cause of belief revision need not be an utterance, but could be any occurrence that provides the agents with some information about the world, such as switching on the lights, observing the outcome of an experiment, or even the failure of some event to occur. Nor do the agents need to be in a face-to-face position: they could, for example, be processes in a distributed system with synchrononous broadcast as the communication medium.

Thus, we will not go into any detailed description of the event causing belief revision. What will be important is that each agent has a specific way of revising its beliefs when this event occurs. Intuitively, this individual revision method is determined by the informational content that the event has for the agent. It represents how the actuating event causes the agent revise its beliefs *about the situation in which the event occurred.* The objective of the present section is to describe how these individual revision methods determine the agents' mutual belief revision.

We now set out to formalize these ideas, stating along the way a number of the assumptions implicit in the formalization. These assumptions are meant to motivate the theory only - they are not intended to entail the formalization.

> **Assumption I.** Agents have perfect introspection, but may be inconsistent.

This assumption may be made precise by taking the logic of agents to be $K45_n$. We showed in the previous section that labelled trees are a canonical class of $K45_n$ structures. Accordingly, we suppose that the class of situations of interest is represented by the set $\mathcal{T} = \mathcal{T}(L, n)$ of labelled trees, where we fix the set L of labels, its associated valuation V, and the number of agents n. The use of $K45_n$ semantic structures does not prevent the theory being applicable in situations where a stronger logic is appropriate as a model of the agents' reasoning powers. All that is required for such an application is to restrict attention to the appropriate subset of trees. We will present an application in the next section based on the logic $S5_n$, which models knowledge rather than belief.

> **Assumption II.** The agents revise their beliefs synchronously, in response to an event ϵ whose occurrence is common knowledge.

If the agents are in an initial state of belief at time t_0, then the assumption of synchronicity means that the belief revision process is complete by some time $t_1 > t_0$, and this fact is common knowledge. In particular, all agents know at time t_1 that the others have completed revising their beliefs.

Let $g : \mathcal{T} \rightarrow \mathcal{T}$ be an operator on $\mathcal{T}$ that represents the relationship between the agents' initial and final states of belief. This operator is specific to the event ϵ under consideration: the mutual belief revision caused by different actuating events will be represented by different operators. Suppose that the initial situation, before occurrence of the event, is represented by a labelled tree $T \in \mathcal{T}$, which describes the facts about the world at time t_0, as well as the agents' mutual beliefs about these facts at time t_0. Then the tree $g(T)$ represents the situation resulting at time t_1 when the agents mutually revise their beliefs because of the occurrence of event ϵ. That is, $g(T)$ represents the facts about the world at time t_1, and the agents' mutual beliefs about these facts at time t_1. Thus, in each case the belief operators B_i are to be interpreted as referring to the current time.

> **Assumption III.** The world is static - it is only the beliefs of the agents that change.

Of course, if an event has occurred then something has changed, but we assume that the truth value of the objective propositions of interest to the agents remains invariant during the belief revision episode. Since these objective propositions are represented by the root of a labelled tree, we may capture part of this assumption by requiring that

$$V(root(g(T))) = V(root(T)) \qquad (1)$$

for each tree $T \in \mathcal{T}$.

Next, let us consider the individual belief revision method of the agents. The intuitive interpretation of these methods is that they capture the agents' revision of belief *about time* t_0. Upon observing the event ϵ each agent asks itself "given what I believe about time t_0, what should have been true at that time for it to be possible (or likely) for ϵ to have occurred." This leads the agent to a new set of beliefs about time t_0.

Since we represent a situation as a labelled tree in $\mathcal{T}(L, n)$, agent i's epistemic state is represented by means of an i-objective forest in $\mathcal{F}_i(L, n)$. Thus, the method used by agent i to revise its beliefs may be represented by a function $\rho_i : \mathcal{F}_i(L, n) \to \mathcal{F}_i(L, n)$ which maps a state of belief to a new state of belief. Like the mutual belief revision operator g, these individual revision functions are specific to the event ϵ under consideration: different actuating events will lead agents to revise their beliefs in different ways.

If the forest F represents the agent i's initial beliefs about time t_0, then the forest $\rho_i(F)$ represents the agent's beliefs about time t_0 after the actuating event has occurred. In particular, in the situation T agent i's beliefs are represented by the forest $child_i(T)$. Thus, if the initial situation is T then agent i's revised state of belief about the initial situation is represented by the forest $\rho_i(child_i(T))$.

However, this forest does not represent agent i's state of belief in the final situation. Since the occurrence of the event ϵ is common knowledge, the agent is aware that other agents are also revising their beliefs. While an agent, after performing individual belief revision, will believe its beliefs about other agents' beliefs at time t_0 to be accurate,[3] it knows that the other agents' beliefs will have changed by time t_1. What it seeks is an accurate state of belief not about time t_0, but about time t_1. How should an agent take other agents' belief revision into account? Clearly, if it does not know how the actuating event will be interpreted by other agents, there is little an agent can say about the other agents' beliefs in the final situation. We therefore make the following (strong) assumption.

> **Assumption IV.** Each agent's revision method is common knowledge.

We now make an intuitive leap in the justification of our theory of mutual belief revision: we claim that if assumptions I-IV are common knowledge then not just the individual belief revision functions, but also the mutual belief revision process itself is known to each agent. The results below will provide some support for this step in the justification, but this argument is, admittedly, circular. As remarked above, our theory consists in the formal definitions, not in the informal assumptions.

Since we have represented the mutual belief revision process by the operator g, the assumption that an agent knows that g is the mutual belief revision operator being applied means that it knows for each situation T' that if T' were the initial situation, then $g(T')$ would be the final situation. Since agent i, after individual belief revision, considers each situation T' in $\rho_i(child_i(T))$ to be a possibility for the situation obtaining at time t_0, it should consider $g(T')$ a possibility for the situation obtaining at time t_1. This suggests the following condition on the function g:

$$child_i(g(T)) = \{g(T') \mid T' \in \rho_i(child_i(T))\} \qquad (2)$$

for all trees $T \in \mathcal{T}$. Intuitively, equation (2) states that the mutual belief revision process operates in two steps: first the agents revise their beliefs about time t_0, then they propagate the resulting state of belief to apply to time t_1, by reasoning about how other agents are modifying their beliefs.

If g is an operator satisfying equations (1) and (2) for all trees $T \in \mathcal{T}$ and all agents i, then we say that g is *compatible* with the revision functions $\rho_1 \ldots \rho_n$. In addition to compatibility, we add some further constraints arising from the slack in our semantic structures. As we argued in Section 2, it is possible for distinct trees to represent what is intuitively the same situation, and for distinct forests to represent what is intuitively the same state of belief. The remedy proposed was the notion of congruence. This motivates the following definitions. Say that the revision function ρ_i is *proper* if $F_1 \cong F_2$ implies that $\rho_i(F_1) \cong \rho_i(F_2)$ for all i-objective forests F_1, F_2. Similarly, say that the operator g is proper if $T_1 \cong T_2$ implies $g(T_1) \cong g(T_2)$ for all trees T_1, T_2. We add to the compatibility condition the requirement that both the revision functions and the operator g be proper.

Do there exist any operators satisfying all these conditions? Note that together with equation (1), equation (2) yields a complete description of the tree $g(T)$, since a tree is determined by its root and its children. Of course, this description is circular, since the children are described in terms of the operator g itself. It is therefore tempting to view equations (1) and (2) as a recursive definition of g. Unfortunately, the standard fixpoint semantics for recursion yields an unsatisfactory answer in this case, because labelled trees are not in general finite, so the recursion never bottoms out and g remains undefined on infinite trees.

However, these intuitions do lead us to a solution. Given the revision functions $\rho_1, \ldots, \rho_n$, define the operator $g = g[\rho_1, \ldots, \rho_n]$ on $\mathcal{T}$ as follows. For each tree $T \in \mathcal{T}$ let $g(T)$ be the set of (L, n)-sequences $l_0 i_1 l_1 i_2 \ldots i_m l_m$ such that there exists a sequence of trees $T_0 T_1, \ldots T_m$ with $T_0 = T$ and

1. $T_j \in \rho_{i_j}(child_{i_j}(T_{j-1}))$ for $j = 1 \ldots m$, and
2. $l_j = root(T_j)$ for $j = 0 \ldots m$.

It is clear from the definition that $g(T)$ is closed under prefixes, and contains a unique (L, n)-sequence of

[3] Note that the formula $B_i(B_i\varphi \to \varphi)$ is valid in K45$_n$ structures.

length zero. Thus $g(T)$ is an (L, n)-labelled tree, and g is a mapping from $\mathcal{T}$ to $\mathcal{T}$. The following result shows that g is a candidate for the mutual belief revision operator.

Theorem 3.1: Let $\rho_i : \mathcal{F}_i(L, n) \to \mathcal{F}_i(L, n)$ be a proper revision function for each agent i. Then $g[\rho_1, \ldots, \rho_n]$ is proper, and compatible with $\rho_1 \ldots \rho_n$.

Because of this result, we will call $g[\rho_1, \ldots, \rho_n]$ the *canonical operator compatible with* $\rho_1, \ldots, \rho_n$. In general, this operator is not the only one compatible with the revision functions - others can be constructed by systematically relabelling the vertices of the values $g(T)$, or by identifying isomorphic subtrees. However, these are essentially the only dimensions of flexibility: as the following result shows, there is a unique solution up to congruence.

Theorem 3.2: If g_1 and g_2 are two operators on $\mathcal{T}$ compatible with $\rho_1 \ldots \rho_n$ then $g_1(T) \cong g_2(T)$ for all $T \in \mathcal{T}$.

Our claim about mutual belief revision, then, is that the mutual belief revision operator g applied when an event occurs should be the canonical operator compatible with the functions ρ_i used by the agents to revise their beliefs about the situation in which the event occurred.

Is this a reasonable account of mutual belief revision? Theorem 3.1 and Theorem 3.2 show that, up to congruence, the canonical operator is the *only* operator compatible with the revision functions. Thus, our theory is reasonable to the extent that it is reasonable to require that mutual belief revision be compatible with individual belief revision. We have informally argued for such a requirement by motivating it from Assumptions I-IV. However, we have not provided a rigorous proof that compatibility is entailed by these assumptions. It may be possible to close this gap by means of a more detailed formalization of the assumptions, but we will not pursue this course here. Instead, in the remainder of the paper we will seek to garner some further support for our claims by presenting two applications of the theory. We will argue that it yields the appropriate answers in both cases. These applications will also serve to illustrate the theory and should help to clarify the intended interpretation of the individual belief revision functions.

4 Scientists in Conference

As a first application of our theory we consider its consequences when the operators B_i are interpreted to refer not to belief, but to the stronger notion of knowledge. There is an independently motivated theory of the dynamics of knowledge in distributed systems [Halpern and Moses, 1990]. This theory gives a very natural semantics to knowledge operators, and has the benefit that it is not necessary to give a separate account of the revision of knowledge. Instead, such an account follows directly from the semantics. In this section we introduce a particular distributed system, and show that states of knowledge in this system are revised in a way corresponding precisely to the theory of the previous section. This gives strong support to the claim that our theory is a reasonable one.

The system we consider is an example of the notion of *knowledge based protocol* [Halpern and Fagin, 1989], in which actions taken depend on the state of knowledge of the agents. In particular, it is a variant of the *communicating scientists* model of [Fagin *et al.*, 1994; Fagin and Vardi, 1986]. In this model a group of scientists, each member of which has made an initial observation of the world, communicate by sending messages in a synchronous system. The world remains static during the communication. We will adapt model this by changing the communication medium from message passing to synchronous broadcast, since this better fits the intended applications of our theory. Accordingly, we refer to our variant of the communicating scientists as the *scientists in conference* model.

We suppose that there are n scientists, and that Φ is a set of atomic propositions, representing objective facts about the world. A state of the world will be an assignment α on Φ. An *initial state* s will be a tuple $(\alpha, S_1, \ldots, S_n)$ where α is an assignment and, for each scientist i, S_i is a set of assignments, intended to represent the set of world states consistent with scientist i's observation. We assume that the observations are accurate, so we require that $\alpha \in S_i$ for each $i = 1 \ldots n$. We also suppose that the scientists have a complete and correct theory of their instruments, so that they know what observations are possible in each state of the world. Semantically, we take I to be a set of initial states, such that an initial state $s = (\alpha, S_1, \ldots, S_n)$ is in I just when the observations S_i are in accordance with the theory, given that the state of the world is described by the assignment α.

Given the set I, we construct the $\mathrm{K45}_n$ Kripke structure $\mathrm{SC}_n(I) = \langle W, R_1, \ldots, R_n, V \rangle$ representing n scientists in conference with initial observations described by I. The worlds, also called *runs*, of this structure will be certain sequences of the form $s \cdot \sigma$, where s is an initial state and σ is a sequence of tuples of the form (i, φ), in which i is an scientist and φ is a formula in $\mathcal{L}_n$. Intuitively, in such a run s describes the objective facts about the state of the world and the scientists' observations of this state, and σ represents a sequence of assertions made by the scientists during the conference. The world is assumed to be static, so the truth value of the atomic propositions at $s \cdot \sigma$ is determined by the assignment in s. Formally, we define the valuation V by letting $V(s \cdot \sigma)$ be the assignment α of the initial state s.

For each scientist i, let the equivalence relation R_i^0 on initial states be given by

$$(\alpha, S_1, \ldots, S_n) R_i^0 (\alpha', S_1', \ldots, S_n')$$

if and only if $S_i = S_i'$. Intuitively, if he makes the same observation in two initial states, the scientist cannot, without obtaining further information, distinguish these states. This relation may be extended to the equivalence relation R_i on the runs of $\mathrm{SC}_n(I)$ by letting $(s \cdot \sigma) R_i (t \cdot \tau)$ just when $s R_i^0 t$ and $\sigma = \tau$. Intuitively, this definition reflects the fact that the scientist is able to observe (and recall) all the assertions made during the conference.

This completes the description of $\mathrm{SC}_n(I)$, except that we have not yet described which sequences are included in the set of runs W. Not every sequence of assertions will yield a valid run. We require that the scientists be *honest*: they should assert φ only if they know φ to be true. More precisely, if $s \cdot \sigma$ is a run then $s \cdot \sigma \cdot (i, \varphi)$ is a run if and only if $\mathrm{SC}_n(I), s \cdot \sigma \models B_i \varphi$. This leads to what at first appears to be a circularity in the definition of $\mathrm{SC}_n(I)$: in order to determine a scientist's knowledge we need to know the set of runs, which in turn depends on the scientists' knowledge.

This sort of circularity is characteristic of the notion of knowledge based protocol, and it has a standard resolution. Intuitively, before any communication has taken place the scientists' states of knowledge are determined by their observations. These states of knowledge then determine what assertions are possible in the first round, and this in turn determines the scientists' state of knowledge after the first round. More generally, the scientists' state of knowledge after k rounds of communication determines what assertions are possible in the next round, and this in turn determines the scientists' knowledge after round $k+1$.

To make these intuitions precise, we define the runs W of $\mathrm{SC}_n(I)$ inductively using a sequence of M_k of Kripke structures. The structure M_k will represent the fragment of $\mathrm{SC}_n(I)$ consisting of the runs of length k. The valuation V and accessibility relations R_i of these structures are defined exactly as for $\mathrm{SC}_n(I)$. The structure M_0 has as its set of worlds W_0 the set I of initial states. Given the structure M_k, we then define the structure M_{k+1} by taking its set of worlds W_{k+1} to contain just those sequences $s \cdot \sigma \cdot (i, \varphi)$ such that $s \cdot \sigma$ is a world of M_k, and $M_k, s \cdot \sigma \models B_i \varphi$. This completes the description of the structures M_k.

The definition of the structure $\mathrm{SC}_n(I)$ is now completed by taking its runs to consist of the worlds $s \cdot \sigma$ contained in W_k for some $k \geq 0$. Because the R_i are defined to be equivalence relations, the structure $\mathrm{SC}_n(I)$ is in fact an $\mathrm{S5}_n$ structure [Halpern and Moses, 1992], so the modal operator B_i may be interpreted as referring to the *knowledge* of scientist i.

Example 4.1: Mathematician P knows the product $N1 \times N2$ of two numbers $N1, N2$ in the range 2 to 100; mathematician S knows the sum $N1 + N2$. They have the following conversation:

S: I don't know what the numbers are, but I know that you don't know either.

P: Now I know the numbers!

S: Aha! Now I also know them!.

We leave it to the reader to verify that this well known example [Gardner, 1979] can be formalized as a run in a certain model of the form $\mathrm{SC}_2(I)$.

Using the unravelling construction of Proposition 2.3, we may associate with any run $s \cdot \sigma$ of $\mathrm{SC}_n(I)$ a tree $T(\mathrm{SC}_n(I), s \cdot \sigma)$ over the set of labels consisting of all runs of $\mathrm{SC}_n(I)$. However, a consideration of the nature of the accessibility relations in $\mathrm{SC}_n(I)$ shows that this set of labels may be simplified. It is readily seen from the definition of the relations R_i that if $t \cdot \tau$ labels a vertex of the tree $T(\mathrm{SC}_n(I), s \cdot \sigma)$ then $\tau = \sigma$, hence all but the initial state t is redundant. This means that the tree associated with a given run may be represented using the set of labels consisting of the initial states I only. Given a run $w = s \cdot \sigma$, define $rep(w)$ to be the (I, n)-labelled tree consisting of the (I, n)-sequences $s_0 i_1 s_1 \ldots i_m s_m$ such that

$$(s_0 \cdot \sigma) i_1 (s_1 \cdot \sigma) \ldots i_m (s_m \cdot \sigma)$$

is in $T(\mathrm{SC}_n(I), s \cdot \sigma)$. Clearly $rep(w)$ is an (I, n)-labelled tree. The considerations above then yield the following result.

Lemma 4.2: For all runs w of $\mathrm{SC}_n(I)$, the tree $rep(w)$ is isomorphic to the tree $T(\mathrm{SC}_n(I), w)$.

This result justifies the use of the tree $rep(w)$ to represent the situation $(\mathrm{SC}_n(I), w)$. We now establish that the theory of belief revision of the previous section can express in a natural way the knowledge revision of scientists in $\mathrm{SC}_n(I)$. We represent situations as trees in $\mathcal{T}(I, n)$. Consequently, a state of belief for scientist i corresponds to a forest in $\mathcal{F}_i(I, n)$, and we require for each scientist a belief revision function $\rho_i : \mathcal{F}_i(I, n) \rightarrow \mathcal{F}_i(I, n)$.

Suppose we wish to represent the belief revision process when scientist S, the *speaker*, asserts the formula φ in $\mathcal{L}_n$. Recall that the intended interpretation of the belief revision functions is that they capture the agents' revision of belief about the initial situation. What information does a scientist gain about the situation corresponding to $s \cdot \sigma$ when the speaker asserts in this situation the formula φ? One way to characterize this information is as being that the run $s \cdot \sigma$

may be extended by the assertion (S, φ). But this is possible just when $\mathrm{SC}_n(I), s \cdot \sigma \models B_S\varphi$. Intuitively, the information gained is that the speaker knows φ.

Thus, when $i \neq S$, we define $\rho_i(F) = \{T \in F \mid T \models B_S\varphi\}$ for all i-objective forests F. Intuitively, this simply adds to scientist i's knowledge about the initial situation the fact that the speaker knows φ. We could define the revision function similarly for the speaker, but note that when the speaker asserts φ, he already knows that he knows φ. Hence when $i = S$, we put $\rho_i(F) = F$ for all $F \in \mathcal{F}_i(I, n)$. It follows from Proposition 2.5(1) that these revision functions are proper.

Our theory of mutual belief revision now states that the mutual belief revision process is represented by $g = g[\rho_1, \ldots, \rho_n]$, the canonical operator on $\mathcal{T}(I, n)$ compatible with $\rho_1, \ldots, \rho_n$. Thus, if the initial situation, corresponding to the run w, is represented by the tree $rep(w)$, our theory predicts that after mutual belief revision actuated by the assertion of φ by S, the scientists' epistemic state is represented by $g(rep(w))$. How does this compare with the revision of knowledge obtained directly from the structure $\mathrm{SC}_n(I)$ itself, which predicts the state $rep(w \cdot (i_S, \varphi))$? The following result shows these predictions to be identical.

Theorem 4.3: If w is a run of $\mathrm{SC}_n(I)$ such that $\mathrm{SC}_n(I), w \models B_S\varphi$, then $rep(w \cdot (S, \varphi)) = g(rep(w))$.

Thus, the theory of mutual belief revision of the previous section accurately describes the mutual *knowledge* revision process of the scientists in conference. Moreover, note that the choice of revision functions needed to obtain this result is very natural. We take this to be strong evidence that our theory is an appropriate model for more the general problem of revision of mutual *belief*.

5 Speech Act Semantics

Speech acts [Austin, 1962] are linguistic actions such as assertions, requests and promises, having an explicit propositional content. One strand in the theory of speech acts, [Allen and Perrault, 1980; Cohen and Perrault, 1979; Cohen and Levesque, 1985; Cohen and Levesque, 1990; Perrault, 1990], seeks to characterize the semantics of a speech act by describing the effect it has on the mental states of the participants of the conversation. In this section we apply our theory of mutual belief revision to give a semantic reconstruction of one such theory, that of [Perrault, 1990], for the simplest type of speech act, assertions.

Perrault observes that the beliefs of the participants after an utterance depend in a complicated way on their beliefs before the utterance. For example, if the speaker sincerely asserts a proposition p and the hearer has no beliefs that would lead him[4] to doubt this assertion, then the hearer should come to believe p also. However, if the speaker lied, or was mistaken, in uttering p, and the hearer knew that p was false when the utterance was made, the hearer does not come to believe p, although he may come to believe that the speaker believes p. Even this latter conclusion does not hold in the situation where it is mutually believed that p is false, and the speaker utters p ironically.

The difficulties pointed out by Perrault may be characterized by saying that in describing how a speech act changes states of belief one faces a (particularly severe) form of the *qualification problem.* Even in simpler domains not involving epistemic states it is often difficult to give a complete set of necessary and sufficient conditions for an action to bring about a given state of affairs. The solution advocated by Perrault is one that has also been proposed as an approach to the qualification problem in other contexts: the use of nonmonotonic logic.

In particular, Perrault uses a modal extension of Reiter's [Reiter, 1980] default logic. The most important operators in this extension[5] are the time-stamped belief operators $B_{x,t}$, referring to the belief of agent x at time t. Each of these operators is assumed to satisfy the axioms and inference rules of the logic KD45 [Chellas, 1980]. Furthermore, these operators satisfy the two axioms

$$\vdash B_{x,t}\,\varphi \supset B_{x,t+1} B_{x,t}\,\varphi \qquad \text{Memory}$$
$$\vdash B_{x,t+1} B_{x,t}\,\varphi \supset B_{x,t+1}\,\varphi \qquad \text{Persistence}$$

whose combined effect is that agents maintain their objective and positive beliefs. There is also a construct $Do_{x,t}(p)$, which represents that agent x utters the proposition p at time t, and an action of observation.

Perrault assumes two basic default rules. One of these is the *declarative rule*[6] $Do_{x,t}(p) \Rightarrow B_{x,t}\,p$, which authorizes the conclusion, if it is consistent to do so, that the speaker believes the proposition she utters. Secondly, there is the *belief transfer rule* $B_{x,t} B_{y,t}\,\varphi \Rightarrow B_{x,t}\,\varphi$, which allows an agent to adopt another agent's beliefs. In order to make the applicability of these two default rules common knowledge, there is also a closure condition, stating that for all agents x and all times t, if $\varphi \Rightarrow \psi$ is a default rule, then so is $B_{x,t}\,\varphi \Rightarrow B_{x,t}\,\psi$. Note that in this system we obtain the following chain of logical and default inferences:

$$\begin{array}{rcl} B_{H,1} Do_{S,0}(p) & \Rightarrow & B_{H,1} B_{S,0} p \\ & \supset & B_{H,1} B_{S,1} p \\ & \Rightarrow & B_{H,1} p \end{array}$$

[4] We will use the masculine pronoun for the hearer and the feminine pronoun for the speaker.

[5] We confine our discussion to the simplest version of the theory: a more elaborate variant also includes operators for intention.

[6] The notation $p \Rightarrow q$ denotes the normal default rule $p : q/q$.

That is, if the hearer believes the speaker to have uttered p then he will first adopt the belief that the speaker believes p, provided this is consistent. If, having done so, it is consistent with his beliefs that p, he will also come to believe p himself.

Perrault works out a number of examples in his theory: sincere assertion, lie, unsuccessful lie, and irony. Here we study the consequences of our theory of mutual belief revision for the first of these examples, the sincere assertion.[7] To do so, we must formulate his assumptions in our framework.

We defer for the moment the question of the appropriate set of trees, and begin with a specification of the revision functions applied when the speaker S asserts p. For the hearer H we define $\rho_H(F)$ to be

$$\begin{array}{ll} \{T \in F \mid T \models p \wedge B_S p\} & \text{if this set} \neq \emptyset\text{, else} \\ \{T \in F \mid T \models B_S p\} & \text{if this set} \neq \emptyset\text{, and} \\ F & \text{otherwise.} \end{array}$$

for all H-objective forests F. Intuitively, this revision function captures the two-step operation of the defaults noted above. Observe that in each case $\rho_H(F)$ is a subforest of the forest F. This reflects the effect of the persistence and memory axioms.

Perrault's default rules are symmetrical with respect to the speaker and hearer, which suggests that we define an identical revision function for the speaker. However, this aspect of Perrault's theory has been criticised by Appelt and Konolige [Appelt and Konolige, 1988]. They point out that this symmetry implies that if the speaker originally does not believe p, then she will become convinced that p is true merely because she asserted it! Since this is counter-intuitive (albeit not unheard of), we assume instead that the speaker's beliefs are not changed by making an utterance, and put $\rho_S(F) = F$ for all S-objective forests F. There is a close correspondence between these revision functions and those we used for the scientists in conference, and indeed our work in this section in the present section may be shown to generalize that of the previous section.

As the next step in formulating Perrault's examples in our model, we need to choose a tree to represent the initial situation. It is not immediately clear what this tree should be, because there is a difference in kind between Perrault's theory and ours. Perrault views his defaults as being applied by an external observer who is reasoning about the beliefs of the participants of the conversation, rather than being applied by the agents themselves. Thus, the formulae in his logic represent the observer's knowledge of the beliefs of the agents. This knowledge, he allows, may be incomplete. But if it is the case that the observer does not know whether an agent x believes φ, what justifies his application of the rule $B_{x,t} B_{y,t}\, \varphi \Rightarrow B_{x,t}\, \varphi$, which surely is meant to reflect a default inference performed by agent x when φ is consistent with the latter's beliefs.

One plausible construal of Perrault's approach that would make the observer's application of this rule valid is to interpret the observer as implicitly applying some additional defaults: if he does not know that an agent believes φ, then he assumes that the agent does not believe φ. In other words, the observer presumes he has complete knowledge of the agent's beliefs. This suggests an *only knowing* formulation [Levesque, 1990] of the initial states of belief of the agents. A number of accounts of only knowing in a multiple agent setting have recently been developed [Halpern, 1993; Lakemeyer, 1993]. The account we sketch here is closely related to these works: we will elaborate on this in the full paper.

Say that a function $f : \mathcal{L}_n \to \{0, 1\}$ is a *satisfiable truth assignment* on $\mathcal{L}_n$ if there exists a K45_n structure M with a world w such that for all formulae $\varphi \in \mathcal{L}_n$ we have $f(\varphi) = 1$ if and only if $M, w \models \varphi$. Take the set of labels L to be the set consisting of all satisfiable truth assignments on $\mathcal{L}_n$, with the associated valuation V defined by $V(f)(p) = f(p)$. We will work with the set of labelled trees $\mathcal{T}(L, n)$. If f is a satisfiable truth assignment, we define the tree $\tau(f)$ to consist of all n-sequences $f_0 i_1 f_1 \ldots i_m f_m$ such that $f_0 = f$ and for all $j = 1 \ldots m$ and all $\varphi \in \mathcal{L}_n$, if $f_{j-1}(B_{i_j}\varphi) = 1$ then $f_j(\varphi) = 1$. The *i-objectification* $\tau_i(f)$ of $\tau(f)$ is defined to be the the tree obtained from $\tau(f)$ by removing all n-sequences with $i_1 = i$.

Proposition 5.1: For all $f \in L$, and all $\varphi \in \mathcal{L}_n$ satisfying $f(\varphi) = 1$, we have $\tau(f) \models \varphi$.

Suppose we are given a tuple $(\Gamma_1, \ldots, \Gamma_n)$ such that Γ_i is a set of i-objective formulae for each $i = 1 \ldots n$. Let f_0 be an arbitrary label. We now define the *only believing tree* $T(\Gamma_1, \ldots, \Gamma_n)$ to be the tree consisting of the n-sequences $f_0 i \cdot \sigma$ for which there exists a label $f \in L$ such that σ is an n-sequence in $\tau_i(f)$ and $f(\varphi) = 1$ for all $\varphi \in \Gamma_i$. This tree is intended to represent a situation in which all that agent i believes is Γ_i, for each $i = 1 \ldots n$. Notice that the i-children of the root are precisely the trees $\tau_i(f)$, where $f(\varphi) = 1$ for all $\varphi \in \Gamma_i$. Our claim that this tree models only believing is supported by the following:

Proposition 5.2: Let φ be an i-objective formula. Then $T(\Gamma_1, \ldots, \Gamma_n) \models B_i\varphi$ if and only if φ is a K45_n logical consequence of Γ_i.

We are now ready to consider the revision of belief in a sincere assertion Specialize to $n = 2$, for agent 1 (the speaker) write S and for agent 2 (the hearer) write H. In the case of a sincere assertion of proposition p, Perrault begins with the theory containing only the for-

[7] The results for the others, to be presented in the final version of this paper, are similar.

mula $B_{S,0}\,p$, together with some formulae representing that S utters p and that the agents are observing each other, so that the fact of utterance becomes common knowledge. Accordingly, we take our initial situation to be the tree $T_0 = T(\{p\}, \emptyset)$, in which the speaker believes only p, and the hearer believes only the vacuous proposition *true*. Letting g be the canonical operator compatible with the belief revision functions ρ_H, ρ_S defined above, our theory predicts the tree $g(T_0)$ as the situation after the speech act.

What do the agents believe in this situation? Perrault argues that there exists a default extension of the original theory that is generated by the original formulae together with the infinite set of formulae

$$B_{S,1}p,\ B_{H,1}p,\ B_{S,1}B_{H,1}p,\ B_{H,1}B_{S,1}p, \ldots$$

He conjectures that this is the only default extension of the theory, and hence an appropriate description of the agent's beliefs after the utterance. That is, according to Perrault, after the utterance it is mutually believed that p. Does the same hold in the tree $g(T_0)$ obtained from our theory? It does not: in fact, even the formula $B_S B_H p$ does not hold in this tree! The reason for this, intuitively, is that the speaker considers it possible that the hearer takes her to be lying, in which case mutual belief is not established by the utterance. More precisely, since all that the speaker believes in T_0 is p, she considers possible a situation T' in which $B_H(\neg p \wedge \neg B_S p)$ holds. In this situation the hearer will not be persuaded to change his beliefs: $\rho_H(\mathit{child}_H(T')) = \mathit{child}_H(T')$. Furthermore, T' has an H-child T'' in which p is false. Thus $g(T'')$ is an H-child of $g(T')$, and since g preserves the valuation at the root we see that $g(T')$ does not satisfy $B_H p$. Because $g(T')$ is an S-child of $g(T_0)$, we obtain that the latter does not satisfy $B_S B_H p$.

If mutual belief of p is not established by the utterance, then what state of belief does our theory predict instead? In order to characterize this state of belief, let ψ be the formula

$$B_H(p \wedge B_S p) \vee B_H(\neg p \wedge B_S p) \vee B_H \neg B_S p$$

Note the correspondence between the disjuncts of this formula and the three cases of the hearer's revision function. Write $M(H, \psi)$ for the infinite set of formulae

$$B_H\psi,\ B_H B_S \psi,\ B_H B_S B_H \psi, \ldots$$

which intuitively corresponds to the hearer believing that ψ is mutually believed. Note that these formulae are S-objective. Similarly, write $M(S, \psi)$ for the corresponding set of H-objective formulae in which the leading modality is B_S. Then following result characterizes the belief state reached after the speech act.

Theorem 5.3:
$$g(T_0) \cong T(\ \{p\} \cup M(H, \psi)\ ,\ \{p, B_S p\} \cup M(S, \psi)\ )$$

That is, in $g(T_0)$, all that the speaker believes is p and that ψ is mutually believed, and all that the hearer believes is p, that the speaker believes p, and that ψ is mutually believed. In particular, it follows from this that after the assertion the formula ψ is mutually believed[8].

Perault's theory predicts that p is mutually believed after a sincere assertion; our theory predicts that it is not. Which is right? We will not offer any answer to this question. Perrault's formalization is itself only a rough approximation to reality, and in modelling it in our framework we made a number of assumptions and modifications. It would be interesting to investigate alternative choices of revision function and initial state to determine precisely under what conditions the proposition asserted does come to be mutually believed. However, we would argue that given the assumptions we have made in the present section, it is very reasonable that the formula ψ should become mutual belief, and that this supports the appropriateness of our account of mutual belief revision.

6 Conclusion

What we hope to have achieved in this paper is the introduction of a semantic framework and to have demonstrated by means of the examples that it is worthy of further study. Much remains to be done. It would be interesting if the circularity in our justification of the framework could be removed. Additionally, we would like to have a general formalism in which the revision functions ρ_i can be defined, and construct for this formalism a calculus that generalizes Theorem 5.3. Such a calculus would facilitate the determination of under what conditions mutual belief is a consequence of a speech act. One could also study the consequences of assuming that the agents' individual belief revision functions satisfy AGM-like postulates. An interesting question would then be to determine how the postulates are reflected in the mutual belief revision process.

Due to space limitations we defer to the full version of the paper a discussion of the related literature, on syntactic accounts of multi-agent belief revision, and on semantic accounts of the revision of knowledge

Acknowledgements

Thanks to Hirofumi Katsuno for comments on earlier versions of this paper. I have also benefited from discussions with Joe Halpern and Gerhard Lakemeyer on only knowing.

[8] It has to be shown that the speaker and hearer both believe ψ. Observe that $B_H(B_H\varphi_1 \vee B_H\varphi_2) \supset (B_H\varphi_1 \vee B_H\varphi_2)$ is valid, which implies that $g(T_0) \models B_S\psi$. That $g(T_0) \models B_H\psi$ follows directly from the fact that $B_H(p \wedge B_S p) \supset B_H\psi$ is valid.

References

[Alchourrón *et al.*, 1985] C.E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[Allen and Perrault, 1980] J. F. Allen and C. R. Perrault. Analyzing intention in utterances. *Artificial Intelligence*, 15:143–178, 1980.

[Appelt and Konolige, 1988] D. Appelt and K. Konolige. A practical nonmonotonic theory for reasoning about speech acts. In *Proc. Annual Meeting of the ACL*, pages 170–178, Chicago, IL, 1988. Association for Computational Linguistics.

[Austin, 1962] J. L. Austin. *How To Do Things with Words*. Oxford University Press, Oxford, 1962.

[Benthem, 1984] J. van Benthem. Correspondence theory. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic II: Extensions of Classical Logic*, pages 167–247. D. Reidel, Dordrecht, 1984.

[Bull and Segerberg, 1984] R. A. Bull and K. Segerberg. Basic modal logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic II: Extensions of Classical Logic*, pages 1–88. D. Reidel, Dordrecht, 1984.

[Chellas, 1980] B. Chellas. *Modal Logic*. Cambridge University Press, Cambridge, 1980.

[Cohen and Levesque, 1985] P. R. Cohen and H. J. Levesque. Speech acts and rationality. In *Proc. 23rd Annual Meeting of the ACL*, pages 49–59, Chicago, IL, 1985. Association for Computational Linguistics.

[Cohen and Levesque, 1990] P. R. Cohen and H. J. Levesque. Rational interactions as the basis for communication. In *Intentions in Communication*. Bradford Boooks, Cambridge, MA, 1990.

[Cohen and Perrault, 1979] P. R. Cohen and C. R. Perrault. Elements of a plan based theory of speech acts. *Cognitive Science*, 3:177–212, 1979.

[Fagin and Vardi, 1986] R. Fagin and M.Y. Vardi. Knowledge and implicit knowledge in a distributed environment. In *Theoretical Aspects of Reasoning about Knowledge: Proc. 1986 Conf.*, pages 187–206. Morgan Kaufmann, 1986.

[Fagin *et al.*, 1994] R. Fagin, J. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, MA, 1994.

[Gärdenfors, 1986] P. Gärdenfors. *Knowledge in Flux: Modelling the Dynamics of Epistemic States*. MIT Press, Cambridge, Mass., 1986.

[Gardner, 1979] M. Gardner. Mathematical games. *Scientific American*, pages 20–24, December 1979.

[Halpern and Fagin, 1989] J. Halpern and R. Fagin. Modelling knowledge and action in distributed systems. *Distributed Computing*, 3(4):159–179, 1989.

[Halpern and Moses, 1990] J. Halpern and Y. Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3):549–587, 1990.

[Halpern and Moses, 1992] J. Halpern and Y. Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54:319–379, 1992.

[Halpern, 1993] J. Halpern. Reasoning about only knowing with many agents. In *Proc. National Conf. on AI*, pages 655–661, 1993.

[Hennessy and Milner, 1985] M. Hennessy and R. Milner. Algebraic laws for non-determinism and concurrency. *Journal of the ACM*, 32(1):137–161, 1985.

[Isozaki and Shoham, 1992] H. Isozaki and Y. Shoham. A mechanism for reasoning about time and belief. In *Proc. Int. Conf. on Fifth Generation Computer Systems*, pages 694–701, 1992.

[Katsuno and Mendelzon, 1991] H. Katsuno and A. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52:263–294, 1991.

[Lakemeyer, 1993] G. Lakemeyer. All they know: A study in multi-agent autoepistemic reasoning. In *Proc. Int. Joint Conf. on AI*, 1993.

[Levesque, 1990] H.J. Levesque. All I know: A study in autoepistemic logic. *Artificial Intelligence*, 42:263–309, 1990.

[Meyden, 1994] R. van der Meyden. Common knowledge and update in finite environments I. In *Proc. 5th Conf. on Theoretical Aspects of Reasoning about Knowledge*, 1994.

[Morgenstern, 1990] L. Morgenstern. A formal theory of multiple agent nonmonotonic reasoning. In *Proc. National Conf. on AI.*, pages 538–544, 1990.

[Moses and Kislev, 1993] Y. Moses and O. Kislev. Knowledge oriented programming. In *Proc. Symp. on Principles of Distributed Computing*, 1993.

[Perrault, 1990] C. R. Perrault. An application of default logic to speech act theory. In *Intentions in Communication*, pages 161–185. Bradford Books, Cambridge, MA, 1990.

[Reiter, 1980] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[Shoham, 1993] Y. Shoham. Agent oriented programming. *Artificial Intelligence*, 60(1):51–92, 1993.

[Thomason, 1990] R. H. Thomason. Propagating epistemic coordination through mutual defaults I. In *Theoretical Aspects of Reasoning about Knowledge: Proc. Third Conf.*, pages 29–39. Morgan Kaufmann, 1990.