The Beliefs of Other Agents

Richmond H. Thomason
Philosophy Department
University of Michigan
Ann Arbor, MI 48109–1003
U.S.A.

`rich@thomason.org`
`http://www.eecs.umich.edu/~rthomaso/`

Version of: April 6, 2001

Abstract

This paper develops a logical account of reasoning about the beliefs of other agents, with the following features.

– It uses multimodal logic to provide a compartmentalization of a single agent's beliefs. The enriched structure that this imposes on the beliefs can be used to show how suitable hypotheses can be formed about the beliefs of other agents.

– It is based on a nonmonotonic logic. This enables plausible inferences concerning the attitudes to be inferred *ceterus paribus*, and provides a natural formalism for representing the exceptional cases.

– It supports formalization of fairly complex, realistic cases of reasoning.

– Under fairly general conditions, it allows mutuality to be supposed.

The technical aspects of this project have been presented elsewhere. This paper, which is intended for philosophers and others who may be interested in the foundational aspects, motivates the underlying ideas and describes some of the applications.

1. Introduction

This paper describes part of an ongoing project. It is designed for philosophers and others who may be interested in foundations of reasoning about attitudes in groups. In [Thomason, 1998], I introduced some of the ideas that I am going to re-present and develop here. That paper described and motivated the underlying logic for intra-agent modality, but did not get much further than this.[1] [Thomason, 2000] motivates a nonmonotonic logic for multi-agent beliefs, concentrating on conditions under which it is possible to obtain *subjective mutuality.*[2]

The purpose of the present paper is to explain and motivate the underlying ideas of the formalism, and to discuss some of the applications. The current version is a working draft: there are some gaps and some rough edges.

The general problem with which I am concerned is how to formalize reasoning about the attitudes of other agents so that we can explain the (apparent) facility of humans at carrying out this sort of reasoning in simple, everyday situations, and can design programs capable of simulating this sort of reasoning.

The problem is challenging. There are several ways to see this; here are two of them.

1. Everyday examples of the following sort show that our beliefs about other people's attitudes are detailed and extensive.

   Case 1.1 *Given:* that a person **a** is sitting next to me on an airplane and is reading an American newspaper. *I believe:* that she believes that Jesse Helms is a senator.

   Case 1.2 *Given:* everything in Case 1.1, and that **a** is an academic. *I believe:* that she doesn't approve of Helms' policies.

   Case 1.3 *Given:* everything in Case 1.2, and that **a** is a philosopher. *I'm not sure:* whether she believes that the frame problem is a problem having to do with reasoning about actions.

   Case 1.4 *Given:* everything in Case 1.3. *I'd guess:* that **a** doesn't know what the qualification problem is, so I'd guess that **a** doesn't believe that the qualification problem has to do with reasoning about actions.

   The level of detail is extremely rich. Any account of the reasoning that isn't equally detailed can't be at all plausible.

   However, our intuitions about the reasoning are relatively shallow. Intuitively, such beliefs seem almost to be immediate; at least, they come to mind more or less effortlessly, and reflecting on them doesn't reveal a breakdown into steps.

   Of course, I can appeal to a rule in each case: for instance, I can say that academics don't approve of ultra-conservative policies, and that Helms has ultra-conservative policies. But this is a restatement of the problem, not a solution. The problem is exactly how we come to have so many detailed rules of this kind regarding attitudes.

[1] Also, some of the details in [Thomason, 1998] concerning the axiomatization of the logic are incorrect.

[2] A proposition p is subjectively mutual for an attitude $\Box$ and group $\mathcal{G}$ if for all $a, a_1, \ldots, a_n \in \mathcal{G}$, $\Box_a[\Box_a p \leftrightarrow \Box_{a_1} \ldots \Box_{a_n} p]$ is true.

2. Fagin et al. argue in [Fagin *et al.*, 1995] that mutual belief is required for many kinds of coordinated activities. Similarly, Clark and Marshall argue in [Clark and Marshall, 1981] that many detailed mutual beliefs are presupposed by conversational activities.[3] If these arguments are correct—and they are rather convincing—we certainly act as if we have many quite detailed and highly contingent mutual beliefs. But the natural way of formalizing the reasoning that would lead to mutual beliefs seems to require arguments with infinitely many steps, each of them highly contingent. Many authors have found it puzzling to explain how we can arrive at mutual beliefs at all.[4]

   As far as I know, there is no adequate formalization of a reasoning process that (1) is at all plausible, in the sense that it does not force patently false assumptions on us, and that (2) can lead to the formation of mutual attitudes in agents. For instance, consider the assumption that any belief acquired by agents **a** and **b** from mutually available evidence in a common situation is mutual. Concede, for the moment, that the assumption is true. Even further, concede that it is mutually believed, without bothering how this mutual belief could have been obtained. Even so, it is difficult to see under what conditions we can apply it to produce mutual beliefs, because of uncertainty over what is mutually available in any realistic situation. Take face-to-face conversation, for example. This is an apparently mutual interaction where evidence is shared. But at any point, an agent's attention can wander, or it can mishear what is said; and the fact that this can happen is itself mutually believed by the agents. Asking for confirmation doesn't resolve the difficulty, since the very same failures can occur during the confirmation process. Therefore, at no point in a conversation will the agents have enough information to apply this rule confidently to infer mutual belief.

2. Shared attitudes in the cognitive and social sciences

2.1. The pervasiveness of agent modeling

Any minimally sophisticated rational community of agents has to presuppose that its members are able to model one another. In particular, in reasoning about their interactions, agents will need a reasonably reliable process of formulating and maintaining hypotheses concerning the attitudes of their peers. Accounts of social interaction presuppose such a mechanism in various ways that are more or less explicit, depending on how their disciplines go about formulating models and theories.

To do even partial justice to the many ways in which attitude modeling is assumed in disciplines such as social psychology, social philosophy, economics, linguistics, and computer

[3] I confine myself to mutual belief, though in discussions of mutual attitudes, authors will often talk about knowledge.

[4] See, for instance, [Parikh, 1990a].

science is far beyond the scope of this paper. Here, I will merely try to illustrate this point with selected comments and references.

Social Psychology. One (admittedly, rather minor) tradition in social psychology investigates shared beliefs.[5] Some social psychologists, at least, are aware that, although it may be convenient to view shared beliefs as collective attitudes of groups, it may be problematic to think of them in this way; see [Cole, 1991] for a discussion of some of the issues.

However, I have found no evidence that social psychologists are much concerned with the foundations of social cognition, or with the mechanisms for forming shared beliefs. Most social psychologists interested in social knowledge concentrate on cases that shade into ideology and on the obvious mechanisms for publicizing, disseminating, and maintaining ideological beliefs as part of a shared culture or group identity. Some social psychologists even seem to feel that there is an inherent conflict between the appropriate methodologies for dealing with individual and social attitudes, so that (apparently) taking cognition as inherently individual would leave no room for social conceptions of knowledge. (See [Bar-Tal and Kruglanski, 1988b].)

Insofar as we want to pursue social psychology as a part of psychology, this perceived conflict must be resolved. To treat shared beliefs as inherently collective would render it impossible to deliver an adequate psychological theory of group attitudes and group reasoning. The burden of proof lies on those who would want to deny this to provide a psychological account of the cognitive mechanisms for social cognition. How, for instance, could a group belief play a role in action that was not mediated by individual cognition? Providing individual cognitive mechanisms that more or less reliably produce mutual beliefs would mitigate the apparent conflict in at least some cases.

Some of the experimental work in social psychology is relevant to my project. Social psychologists have documented cases, such as “false consensus,” in which individuals are regularly misled about the opinions of their peers.[6] These cases show that any mechanism for inferring the beliefs of other agents has to be not only unreliable, but has to be capable of systematic bias. But the cases in which false consensus occurs are ones in which communication is suppressed and, as far as I know, this sort of unreliability has not been documented in normal, face-to-face conversations on topics that are not emotionally charged. I assume that a theory of human reasoning about beliefs of other humans has to provide for the cases in which the reasoning is reliable, but to leave room for the other cases as well. This paper is devoted to the first of these issues.

Developmental Psychology and Folk Theories of Mind. In the psychological and philosophical literature on folk theories of mind, there is a group that champions the idea that we model other agent’s attitudes by putting ourselves in their shoes. The idea has generated a certain amount of controversy (see, for instance, [Davies and Stone, 1995a, Davies and Stone, 1995b]). I am not sure whether I agree with some of the philosophical claims associated with this view, and apparently the psychologists disagree about how well it is supported by the developmental evidence. Despite these doubts about how simulation fits into a general cognitive model of folk psychology, I am sure that simulation is plausible as reasoning mechanism. For instance, in Case (1.1), I form the belief about my seatmate’s be-

[5] See, for instance, [Bar-Tal and Kruglanski, 1988a], [Resnick *et al.*, 1991].

[6] See, for instance, [Marks and Miller, 1987].

liefs by seeing whether I myself have an American-newspaper-reader belief that Jesse Helms is a senator.

The problem is that although simulation is suggestive, it is hardly more than this. The simulation metaphor is compatible with many quite different detailed models of the reasoning, and, as far as I can see, is unconstrained by the experimental evidence.[7] To an introspecting casual observer, simulation may seem like a simple, unanalyzable process, as simple as literally putting on another person's shoes. But any flexible, powerful form of simulation is highly unlikely to be cognitively simple. For instance, consider a universal Turing machine.[8] This is capable of simulating any Turing machine. But just because of this flexibility, a universal Turing machine can't do this simulation by simply *becoming* another machine. To do that, the universal machine would have to have infinitely many states, and so would not be a Turing machine. Instead, it performs the simulation by manipulating a description of the other machine according to the general rules for Turing computation.[9]

Looseness in the model of simulation as a form of reasoning creates a certain looseness of fit in relation to psychological evidence and allows apparently competing simulation theories to proliferate unchecked. I suspect that this looseness of fit may be in part responsible for the controversy in the literature about the cognitive adequacy of simulation theories.[10] According to the logical model I will propose below, simulation is not a matter of an entire agent modeling another entire agent; the simulating agent chooses a specialized cognitive module to model a corresponding part of another agent's cognition. The adequacy of the simulation depends on how aptly the specializing simulator is chosen.

Social Philosophy. This section remains to be written. The point I want to make is similar to the one I tried to make about social psychology. Although some recent philosophers (see especially [Gilbert, 1989], [Tuomela, 1995]) have stressed the importance of shared attitudes and coordinated action, this work leaves the relation of these things to individual cognition unclear.

Economics. The idea of introducting an explicit account of knowledge into game theory is primarily due to Robert Aumann; see [Aumann, 1976]. These applications have been extended to bargaining theory; see, for instance, [Milgrom and Stokey, 1982]. Mutual knowledge of certain background conditions and of the economic rationality of all the participations in a transaction has emerged as a central precondition of many of the fundamental results of microeconomics; see, for instance, the surveys [Geanakoplos, 1990,

[7] Here are some of the important questions that simulation fails to answer. (1) What are the mechanisms for constructing and maintaining systems of belief? (2) In modeling another agent's beliefs, to what extent are similar cases remembered, and to what extent is the model constructed on the bases of *ad hoc information*? (3) Exactly what evidence about other agents do we use in simulating them? (4) How can we simulate another agent's beliefs when many aspects of any other agent's experience are totally inaccessible to us?

[8] See [Turing, 1936], or any textbook on the theory of computation.

[9] The relevance of this point to a psychological theory of simulation is unclear. But most of this unclarity is due to the current state of the psychological theories. In any case, I find the point suggestive.

[10] The main competitor of the simulation theory is the "theory theory," according to which we understand other agents by forming a theory of them. But notice that the universal Turing machine simulates other Turing machines in part because in incorporates a crucial part of the theory of Turing machines. It is perfectly possible to simulate *by means of* a theory. And there are many cases—virtual reality is one—where no one knows how to construct an effective simulation without a highly articulated theory of what is being simulated.

Geanakopolos, 1992]. The foundational centrality of mutuality in this area, I believe, is due more to theoretical sophistication and intensive work on foundations than to the specific subject matter.

Linguistics. Leaving aside work in the philosophy of language inspired by [Lewis, 1969], the most important and sustained linguistic work on shared attitudes has been carried out by Herbert H. Clark and his associates. Clark is a psycholinguist interested in pragmatics. His theoretical work and experiments stresses the importance of *grounding* in discourse; the establishment of shared material which is used as a basis for the planning and interpretation of referring expressions, and for many other reasoning processes in discourse. Clark makes a good case that grounded material must be mutually believed, in the technical sense. See [Clark, 1992a] for a collection of many of the relevant papers. Similar ideas have also appeared in the philosophical literature on pragmatics; see [Stalnaker, 1975, Lewis, 1979, Thomason, 1990]; the philosophical work reinforces the foundational importance of mutuality but does not integrate the ideas with empirical data as Clark does.

Computer Science.

2.2. Need for an account of agent modeling

Normal, adult humans are (apparently) very good at agent modeling. There are limits to our ability to form reliable hypotheses about the beliefs, intentions, and desires of our fellow human beings, but within these limits, and with others who share our background and culture, we can be remarkably flexible and reliable.

A process this flexible, this reliable, and this dependent on the interpretation of evidence, must make use of sophisticated reasoning. For some reason, though, it is difficult to obtain insight into the structure of this reasoning, to break it down into steps and to formalize it, on the basis of introspection alone.

3. Mutuality as a reasoning problem[11]

We now turn from observed interpersonal reasoning about belief to theories of interpersonal reasoning. None of the available theories of reasoning about belief (roughly, these are the logical, computational, and microeconomic theories described in [Fagin *et al.*, 1995]) have much to say about how we obtain and maintain opinions about the beliefs of other agents, but they all assume to some degree that we have such opinions. In particular, many different approaches to interpersonal reasoning have found it important to assume that agents are capable of achieving mutuality. Showing how agents can infer mutuality under realistic circumstances therefore becomes an important test of any account of how agents obtain beliefs about other agents' attitudes.

The problem of mutuality can be stated informally, but I don't believe it is possible to articulate it clearly without some logic; so this is where we will begin.

[11] This section corresponds to parts of [Thomason, 2000].

3.1. Logical background

I use modal logic to formalize propositional attitudes. The tradition of epistemic modal logic goes back to [Hintikka, 1962]. Its usefulness as a tool in formalizing reasoning about propositional attitudes was not realized by the subsequent tradition in philosophical logic—chiefly, I think, because pretty glaring foundational problems are associated with this approach. The most challenging of these problems have to do with hyperintensionality.

Philosophers do not seem to be very good at getting the most out of theories that suffer from such foundational problems. But also the applications to problems in group reasoning were not apparent in the 1960s, and the crucial generalization of epistemic logic to the multiagent case did not seem important to the modal logicians of that time.[12]

Applications of epistemic logic were revived by computer scientists interested in knowledge-based analyses of distributed systems. Joseph Y. Halpern took the lead in this development, organizing a series of conferences bringing together computer scientists, economists, and logicians.[13] These conferences, and the systematic presentation in [Fagin *et al.*, 1995] of what can be done with a multiagent modal logic, make an impressive case for this approach to reasoning about propositional attitudes. The modal approach can help to illuminate important issues. We shouldn't forget the foundational problems, but we need at the same time to recognize that there is useful work to be done by modal theories of belief.

Of course, there are also more general philosophical problems concerning the use of logic for such a purpose. In relation to epistemic logic, these problems do not seem to be significantly different from the problems concerning any other approach to idealized rationality—including any logical theory, probability theory, idealized theories of computation, and microeconomic theories. I see no point in dealing with these problems at a retail level.

I do not mean to dismiss these problems absolutely and permanently, but I do want to suggest that it may be useful and productive to keep them at arm's length.

Hopefully, what has been said up to this point will suffice to motivate the following project. We will work with a modal logic in which the operators are indexed: $\texttt{[}\,i\,\texttt{]}(p)$ means that p is known, or believed, with respect to index i. I prefer to work with belief, but for reasons that I hope will become clear later I think that the term 'belief' has misleading connotations; 'supposition' might be better. Typically, the index i stands for an agent; but I will be using a more complicated indexing scheme. The goal is to create a formalism that could in principle provide for everyday reasoning concerning the beliefs of other agents. Accounting for the achievement of mutuality will provide an initial challenge which will serve to test the theory.

3.2. What is mutuality?

A proposition p is *mutually believed* by a group $\mathcal{G}$ in case every member of the group believes p, and believes that every member believes p, and believes that every member believes that

[12] For my own part I was certainly aware of these generalizations, but they didn't strike me as logically interesting because the interaction between the modalities for multiple agents didn't seem to create a need for any new axioms.

[13] These were the TARK conferences on theoretical aspects of reasoning about knowledge. There were six such conferences, held from 1986 to 1996: [Halpern, 1986], [Vardi, 1988], [Parikh, 1990b], [Moses, 1992], [Fagin, 1994], and [Shoham, 1996].

every member believes p, and so forth.

For theoretical purposes, it seems to suffice to consider only two-member groups. From here on, I will assume that the group $\mathcal{G}$ of agents contains just two members, **a** and **b**.

Some more notation: let $\alpha_i \in \{a, b\}$. Then $[\alpha_1 \ldots \alpha_n] = [\alpha_1] \ldots [\alpha_n]$. Where α is a string over the alphabet $\{a, b\}$, α^n is the string consisting of n repetitions of α. In particular, α^0 is the empty string.

Clearly, mutual attitudes involve iterated attitudes. It will be useful to make this idea of an iteration precise. The following two definitions do this.

Definition 3.1. *Iteration depth.*

1. p $[a]$-iterates to depth 0 for all p.
2. p $[a]$-iterates to depth $1 + 2n$ in a model or example iff $[a(ba)^n]p$ is true in the model or example.
3. p $[a]$-iterates to depth $1 + 2n + 1$ in a model or example iff $[a(ba)^n b]p$ is true in the model or example.
4. No p $[a]$-iterates to depth ζ for any $\zeta \geq \omega$.

Definition 3.2. *Iteration complexity.*

The $[a]$-iteration complexity of p in a model or example is the smallest ordinal $\eta \leq \omega$ such that p does not iterate to depth $\eta + 1$ in the model or example.
$[b]$ iteration complexity is defined analogously to $[a]$ iteration complexity.

In a case where $[a]p$ is false, the $[a]$ iteration complexity of p is 0. If $[a]p$ is true but $[a][b]p$ is false, the $[a]$ iteration complexity of p is 1. If $[a]p$ and $[a][b]p$ are true but $[a][b][a]p$ is false, the $[a]$ iteration complexity of p is 2. If the $[a]$ and the $[b]$ iteration complexity of p are both ω, p is mutual for the group $\{a, b\}$ and the attitude $[\,]$. If the **a**-*believes* and the **b**-*believes* iteration complexity of p is greater than 0, but finite, we have a case in which both **a** and **b** believe p, and they may believe that each other believes p, and so forth, but at some finite point the iteration plays out. It is possible to construct plausible examples where the **a**-*believes* and the **b**-*believes* iteration complexities of p are finite but greater than one. But as the complexity increases, the examples become more contrived and more difficult to think about intuitively. Some authors have noticed that human reasoners are not very good at reasoning about finite levels of interation complexity greater than 2; people seem to be most comfortable with 0, 1, and ω. The formalization that I provide below goes a small way towards explaining this phenomenon.

3.3. Immediacy theories

The most detailed studies with which I am familiar of the reasoning leading to mutual attitudes[14] all suggest that an immediacy-based approach can account for the reasoning. The three accounts are similar, the two later ones being influenced by Lewis' approach.

[14] By David Lewis, Stephen Schiffer, and Herbert Clark and Catherine Marshall. See [Lewis, 1969], [Schiffer, 1972], [Clark and Marshall, 1981].

None of them is fully formalized; they all suggest that mutuality is somehow precipitated by the reflexive character of certain shared situations.[15]

I want to explain why I think that this idea fails to provide an adequate formalization of the relevant reasoning. To make the point, I'll use a simple version of the immediacy theory that doesn't correspond exactly to any earlier presentation of the idea. Although this won't do full justice to any of the views I mean to undermine, I don't think it will leave out anything essential or weaken the force of the arguments.

Let's say that a proposition p *guarantees q-mutuality* for agent **a** and group $\{a, b\}$ in an example or model if

$$\textbf{Mut Ax} \qquad [\,a\,][p \rightarrow [\,\textsc{mut}\,]q]$$

holds in the example or model.

The idea of immediacy is that agents often find themselves in circumstances where something like **Mut Ax** can apparently be used to secure mutuality. The following sorts of examples are used to illustrate the point. In Example 1, (1a) represents p in **Mut Ax** and (1b) represents q.[16] In Example 2, (2a) represents p in **Mut Ax** and (2b) represents q.[17]

(1a) **a** and **b** are sitting across from one another at a small dining table, looking at a candle on the table and at each other looking at the candle.

(1b) There is a candle on the table.

(2a) **b** says to **a**, in a face-to-face conversation, "I will be here tomorrow at 9am." The day of the utterance is d, the place of utterance is l.

(2b) **b** will be at l at 9am on $d + 1$.

In Example 1, what guarantees mutuality for (1b) is the proposition that **a** and **b** are face-to-face across a small table, and are both looking at the candle—or maybe a qualification of this proposition.[18] In Example 2, what guarantees mutuality for (2b) is the proposition that, in a face-to-face conversation, **a** has just said to **b** "I will be here tomorrow at 9am." I don't want to quarrel with the insight behind these examples: that circumstances such as these allow mutuality to be inferred. In fact, I want to say that in these circumstances, each agent can infer an infinite iteration complexity for the appropriate propositions. But I don't want to say that an axiom like **Mut Ax** *guarantees* this inference.

My explanation of why I think that immediacy theories provide an inadequate account of the reasoning is divided into three topics: incrementality, monolithicity, and fallibility.

[15] It is also true that there are differences in the three versions, and that these could lead to different formalizations; and [Barwise, 1988] suggests different formalizations of [Lewis, 1969] and [Clark and Marshall, 1981]. However, I don't believe that these differences are relevant to the points I wish to make.

[16] The example is from [Lewis, 1969, pp. 52–57].

[17] An example from [Schiffer, 1972, pp. 31–36].

[18] Schiffer produces a fairly elaborate qualification, [Schiffer, 1972, p. 35].

3.3.1. Incrementality issues

Finite iteration complexity can arise naturally. A classic series of examples of increasing complexity is presented in [Clark and Marshall, 1981]; another is developed in Section 9, below, where for a certain p, $\texttt{[}a\texttt{]}\texttt{[}b\texttt{]}p$ holds but $\texttt{[}a\texttt{]}\texttt{[}b\texttt{]}\texttt{[}a\texttt{]}p$ does not. By making inferences of mutuality take place in a single step whenever they do take place, immediacy approaches allow the reasoning that produces to mutuality to fail in only one way—by blocking all iteration depths. A reasoning process that leads to an iteration complexity of, say, 2, apparently must involve entirely other forms of reasoning. An account of the reasoning that is more unified than this, I think, would be more plausible and more explanatory.

3.3.2. Monolithicity issues

Mut Ax implies $\texttt{[}a\texttt{]}p \rightarrow \texttt{[}a\texttt{]}\texttt{[}b\texttt{]}p$ and $\texttt{[}a\texttt{]}p \rightarrow \texttt{[}a\texttt{]}\texttt{[}b\texttt{]}\texttt{[}a\texttt{]}p$. But this second formula is not plausible in many cases where mutuality is wanted. Suppose that Ann is an attorney defending a client, Bob, whose story she does not believe. In interviewing her client, she suspends her disbelief; professionally, she has to take his story at face value. Her interview with Bob makes use of many presuppositional features of conversation that are generally assumed to require mutuality—features like definite reference. She could say, "Now, after you got up from your nap, did you make any phone calls?" But she doesn't in fact believe that Bob took a nap while the robbery he is accused of took place. Bob doesn't believe that she believes this. But the interview takes place without any presuppositional anomalies, without any of the abnormalities that accompany failures of mutuality.

Robert Stalnaker has discussed similar problems ([Stalnaker, 1975]), and has proposed a natural solution: invoke an attitude of "belief for the sake of conversation." Ann and Bob are modeling not each other's beliefs, but each other's C-suppositions, suppositions for the sake of this particular conversation. For instance, $\texttt{[}a\texttt{]}A \rightarrow \texttt{[}a\texttt{]}\texttt{[}b\texttt{]}\texttt{[}a\texttt{]}p$ represents Ann's C-supposition that Bob C-supposes that Ann C-supposes that p. Their mutuality is possible because the conversants are constructing, for this conversation, a special purpose attitude that not only serves to keep track of the conversation but that maintains mutuality. Under some circumstances, this local mutuality may precipitate actual mutual belief, but these circumstances are decoupled from the rules that govern conversation.[19]

The achievement of mutuality in conversation, then, depends on the ability of the participants to construct at the beginning of the conversation an appropriate *ad hoc* attitude, which from one point of view models the content of the conversation, and from another point of view models for each participant the other participants' views of this content.

Now, the things that are supposed for the sake of conversation will include not only what has been contributed to the conversation by speech acts, but what the participants can reasonably expect to be mutual at the outset.[20] We are therefore assuming that agents must be able to associate observable properties of other agents with an appropriate initial attitude which is assumed to be mutual. And if mutuality is taken to be a mark of successful conversation, then we are also supposing that agents often initialize C-supposition in much

[19] The work done by the distinction between C-supposition and belief is similar the work done by J.L. Austin's distinction between illocutionary and perlocutionary acts.

[20] See [Clark and Schober, 1989, pp. 257–158].

the same way.

Suppose, for instance, that Ann meets Bob at an AAAI conference, and begins a conversation. This supposition activates a large number of specific hypotheses about beliefs she presumes to be mutual, which can be used as a basis for the conversation. I propose to explain how these hypotheses are generated by assuming that Ann's organization of her own beliefs is not monolithic, that she has indexed her beliefs in such a way that she can quickly access prominent beliefs that are common to computer scientists.

By introducing special-purpose modalities whose purpose is mutual modeling into the makeup of single agents, we have a more plausible way of producing iterated attitudes. I will elaborate this idea below, in Section 8, and will show how this idea can lead to infinite iteration complexities. Under favorable conditions, at least, this will produce full mutuality.

The technique of modeling a single agent's beliefs by a family of modal operators is not needed to account for many of the logical issues involved in mutuality. But it is indispensible to account in practice for a wide range of examples that involve mutuality.

3.3.3. Fallibility

The examples like (1) and (2) above that are usually given to motivate immediacy theories of mutuality are chosen to conceal the worst flaw of these theories—the defeasibility of the associated reasoning. If you and I are sitting at a small table, looking at a candle on the table and at each other looking at the candle, it is hard to see how we could fail to mutually believe there is a candle on the table. But this is due to some extent to the fact that, when an example is sketched, we tend to imagine a normal case; you might well not realize you are looking at a candle if it is a novelty item designed to look like a wine bottle. We often need to deal with cases that are not so straightforward. If a group of five quarters is lying on the table, I'm pretty safe in assuming that we mutually believe there is $1.25 on the table, but I could well be wrong. If there are eleven quarters, the assumption that we mutually believe there is $2.75 on the table is riskier, but I might well make it.

The assumptions we make in initializing a conversation that are not based on immediately perceived mutual situations are even more patently defeasible. Part of being a skilled conversationalist is to make such assumptions, realizing at the same time that they may be incorrect, and having a notion of how to correct things if the assumptions should fail. Maybe Bob is a book exhibitor rather than a computer scientist. Once Ann finds this out, she will probably have an idea of where the conversation went wrong, and how to adjust it. Even a conversation that begins with the participants fully coordinated can lose mutuality, because of ambiguity and inattention. Skilled conversationalists are able to identify and repair such failures.[21]

Axioms like **Mut Ax** are inadequate in accounting for such phenomena. To qualify as an axiom, **Mut Ax** has to hold across all the examples in its intended domain of interpretation. In most cases when we want to imagine that interacting agents apply **Mut Ax**, the agents would be well aware that the axiom could be false, and so that it lacks the properties of an axiom. We could try to remedy this difficulty by using refined axioms of the form

Qualified Mut Ax $\quad [a][[p \wedge r] \rightarrow [\mathit{mut}]q]$,

[21] See [Mortensen, 1996].

where r is is a conjunction of clauses which together eliminate the cases in which mutuality could fail.

But this merely relocates the problem. First, though it is often possible to find reasonable qualifying conditions, and to further improve these by further refinements, it seems impossible in realistic cases to bring the process of refinement to an end.[22] Second, we are typically willing to use **Qualified Mut Ax** without explicitly checking the qualifying conditions. These two circumstances are best dealt with by using a nonmonotonic logic.

The approach that I develop in this paper combines solutions to all three of these problems. I use a nonmonotonic logic, which secures ω-level iteration complexity in one step in the normal cases, but which in principle could fail at any finite iteration level. Within this logic, it is possible to develop a theory of exceptions to the normal case. Such a theory, I believe, is an essential part of any solution to the problem of inferring mutuality, since this reasoning is failure-prone, and agents need informed ways of recovering from failures. Finally, the logic is based on the intra-agent modality of [Thomason, 1998] and so provides an approach to the problem of initializing mutual attitudes.

4. Subagent simulation as an agent modeling mechanism

Assuming that we are committed to the modal model of belief, then the natural way for an agent to represent the belief of another agent would be to construct a modal operator: a model's b's beliefs by constructing a modal operator $\texttt{[}a, b\texttt{]}$. (We need two indices here because this operator is a's representation of b's attitude.)

But if we think of things in this way, the modeling task is plainly impossible. Every agent b will have many private experiences and memories that another agent a can't possibly hope to guess at. Moreover, there would be no point in guessing at most of these beliefs. In any information transaction, a will be interested in a very limited part of b's repertoire; it would not only be impossible, but be beside the point for a to form a hypothesis about the whole of b's beliefs.

Here, an idea that first appeared, as far as I know, in [Stalnaker, 1975] is useful. The participants in a conversation mutually construct a model not of each others' beliefs, but of what is *supposed for the sake of the conversation*. To put it in the terms of [Clark and Marshall, 1981] (which apparently was not influenced by Stalnaker) the participants construct only the *common ground*. Or, in the terminology of [Thomason, 1990], they construct the *conversational record*.

This is a much more feasible task that constructing the whole of another agents' beliefs. In fact, I believe that we can understand a great deal about the workings of conversation by assuming that it is designed to facilitate this task. However, even this simplified reasoning task is far from simple, and very little has been done to model it in detail. One of the purposes of this paper is to fill in this gap.

The idea is now that in the course of a conversation, or other transaction in which it is important to model part of another's attitudes, an agent a constructs an attitude $\texttt{[}a, b, i\texttt{]}$, where i is an index representing the relevant part of b's beliefs.

To illustrate how this might occur, we return to the example of Ann and Bob. Ann has

[22] This is the qualification problem. See, for instance, [McCarthy, 1977, Thielscher, 1996].

kept track of the things she learned in becoming a computer scientist. She expects computer scientists to have organized their beliefs in much the same way. For instance, she not only expects Bob to have learned

> p: finite state automata accept regular languages

but that any computer scientist can be expected to have learned this. A modal model of Ann's mental contents will therefore contain not only $[a, \mathrm{CS}]p$, but $[a, \mathrm{CS}][b, \mathrm{CS}]p$.

Now, Ann has kept track of many other beliefs in the same way. In fact, each time she learned something, she associated a set of indices with the new belief, each index corresponding to a modal operator. For instance, she has such an index for English-speaking Americans, and for people attending the conference at which she finds herself. This makes it possible for her to construct a special-purpose modality for this conversation, based on the propositions that she would expect everyone who is a computer scientist, an English-speaking American, and a conference attendee to suppose.

This provides a way of implementing the proposal from [Clark and Schober, 1989, pp. 257–158][23] for how conversants acquire a common ground. Their conditions, which are stated in terms of speech communities, can be implemented using the mechanism of indexed modalities, assuming that appropriate indices can be more of less reliably attributed to corresponding components of other agents' beliefs.[24]

> The common ground between two people—here, Alan and Barbara—can be divided conceptually into two parts. Their *communal common ground* represents all the knowledge, beliefs, and assumptions they take to be universally held in the communities to which they mutually believe they both belong. Their *personal common ground* represents all the mutual knowledge, beliefs, and assumptions they have inferred from personal experience with each other.
>
> Alan and Barbara belong to many of the same cultural communities ...
>
> 1. *Language*: American English, Dutch, Japanese
> 2. *Nationality*: American, German, Australian
> 3. *Education*: University, high school, grade school
> 4. *Place of Residence*: San Francisco, Edinburgh, Amsterdam ...

This modularization of individual beliefs matches well with the way in which people seem to learn propositions; frequently, if not typically, we can recall not only what we believe, but the circumstances under which we came to have these beliefs. Also, I believe that this architecture would be useful in many other reasoning tasks.[25]

On the other hand, it may be somewhat misleading to speak of "belief" once this step has been taken. Things that we take another agent to believe, or that we suppose for the sake of a conversation, are not necessarily things that we ourselves believe.[26]

[23] Page numbers from the version in *Arenas of Language Use*.

[24] The term "community" is a little misleading. We might well create an index for people who have read the morning newspaper, match it to an interlocutor, and use it in a conversation. But the people who have read the morning newspaper don't constitute a community in any normal sense of the term.

[25] Belief revision is one example.

[26] The logical theory presented below doesn't have the capacity to model this. It could be done by making inheritance of attitudes nonmonotonic, but this is a step that I have not yet taken.

These ideas lead to the following program: the first step in modeling *inter-agent* reasoning about attitudes is to create a theory of *single-agent* attitudes that allows the modeling to take place. Single-agent attitudes need to be indexed to the circumstances in which they were learned. These circumstances should be formulated in a way that facilitates matching them to other agents. Suppose, for example, that an agent **a** acquires a belief p. We include in the representation of the resulting belief an index i standing for certain features of the circumstances under which **a** acquired p. For instance, i could represent high-profile, frequently repeated American newspaper information. If **a** sees **b** reading an American newspaper, **a** can reasonably suppose that **b** believes p. I will call the index i (or, more precisely, the pair $\langle a, i\rangle$) a *subagent*.

From uses of multiple modalities in logical models of multi-agent systems (see [Fagin *et al.*, 1995]) and contextual reasoning (see [Buvač and Mason, 1993]), we are familiar with the idea of modal logics in which indices are attached to the modalities, where these indices stand either for agents or microtheories. I propose to use this apparatus to model the modularization of single-agent belief that is required in the computer science conference example of Section 3.3.2. Ann's beliefs in that example are now to be represented not by a single modality $[a]$, but by a family of modalities $[a, i]$, where $i \in \mathcal{I}_a$. Here, $\mathcal{I}_a$ is a set of "subagents," or indices standing for special-purpose belief modules. In the example, Ann uses a modality $[a, \mathrm{CS}]$ that singles out things that any computer scientist could be expected to have learned.

The general idea is similar to modal theories of context, such as that of [Buvač and Mason, 1993]. For instance, there will be "lifting rules" that govern transfers of information among the subagents of a single agent. Although an agent a can obtain information from another agent b (for instance, by communication), this is not a matter of a's internal epistemic organization, and we certainly do not want to relate indices $\langle a, i\rangle$ and $\langle b, i\rangle$ by lifting rules. But a's beliefs about b's beliefs do in general depend on beliefs of the subagents of a that imitate subagents of b; so we will have lifting rules (that may need to be nonmonotonic), rules that relate beliefs of some of a's subagents to a's beliefs about b's beliefs.

Although the logic is similar to modal logics of context, there are extra complications due to the need to distinguish intra-agent from inter-agent modalities. I begin with the intra-agent logic.

5. Modeling the multiplicity of single-agent beliefs[27]

Some subagents can *access* other subagents. This is not a form of communication; it means that the information available to the accessed subagent is automatically available to the accessed subagent.[28] I will not go into details here, but I believe that this modular organization of the individual's epistemology is useful for the same reasons that make modularity useful in knowledge representation. There are, of course, many analogies between the organization of large-scale knowledge bases into microtheories, as discussed in [Guha, 1991] and the organization of individual attitudes that I am proposing here.

When a subagent i does not access j, I will assume that j is entirely opaque to i. We

[27] Some of this section corresponds to parts of [Thomason, 1998].

[28] I am talking here about subagents of the same agent.

might model this by disallowing formulas like $\texttt{[}i\texttt{]}\texttt{[}j\texttt{]}A$, but linguistic restrictions of this kind are in general less satisfactory than a semantic treatment. So I will assume that $\texttt{[}i\texttt{]}\texttt{[}j\texttt{]}A$ is false if i can't access j.

These ideas lead to the following language and satisfaction condition for modal formulas.

Definition 5.3. *Intra-Agent Modal Languages, Modal Satisfaction.*
An intra-agent propositional language $\mathcal{L} = \langle \mathcal{I}, \preceq, \mathcal{P} \rangle$ is determined by the nonempty set $\mathcal{I}$ of indices, a reflexive, transitive ordering $\preceq$ over $\mathcal{I}$ and a nonempty set $\mathcal{P}$ of basic propositions. Where $\mathcal{M}$ is a model, $\mathcal{M} \models_{i,w} \texttt{[}j\texttt{]}A$ iff $i \preceq j$ and for all $w \in R_j w'$, $\mathcal{M} \models_{j,w'} A$.

I will not go into the details of the logic; these are presented in [Thomason, 1998, Thomason, 2000]. Think of a family of modalities $\texttt{[}i\texttt{]}$, where i is a subagent index. I assume that the relation over possible worlds corresponding to these modalities is Transitive, Euclidean, and Serial; this combination is generally used in contemporary logical models of single-agent belief; see [Fagin *et al.*, 1995].

Departing from the usual practice in modal logic, I assume that $\texttt{[}i\texttt{]}\texttt{[}j\texttt{]}p$ is false if $\not i \preceq j$; this is meant to make a subagent unaware of other subagents that are not accessible to it. The resulting logic is a multimodal version of the non-normal logic **E2** that is formulated in [Lemmon, 1957] and proved complete in [Kripke, 1965]. As far as I know, the non-normal modal logics are usually considered to be exotic and more or less useless. But they appear to be very useful in cases of this sort, in which there is a clear motive for limiting accessibility.

6. Modeling the beliefs of many agents[29]

We now want to imagine a community of agents. Each agent has modularized beliefs along the lines described above. But in addition, each has beliefs about its fellow agents; and these beliefs iterate freely. In fact, for multi-agent beliefs I want to adopt the familiar framework of [Fagin *et al.*, 1995].

Intra-agent and multi-agent epistemic logic are fundamentally different. In the latter case, agents form opinions about other agent's beliefs in much the same way that they form opinions about any other feature of the world. In the former case, when $i \preceq j$, then j represents a part of i's opinion, and i directly accesses j in recalling its opinions.

We will need indices for agents as well as for the associated subagents. Thus, we will have formulas like

$$\texttt{[}a,i\texttt{]}[p \rightarrow \texttt{[}b,j\texttt{]}[q \rightarrow \texttt{[}a,i\texttt{]}r]],$$

where a and b are agent indices. This formula says that **a**'s i-module believes that if p then **b**'s j-module believes that if q then **a**'s i-module believes that r. The notation assumes that the overall modularization of each agent's beliefs is the same.

Definition 6.4. *Inter-Agent Modal Languages.*
An inter-agent propositional language $\mathcal{L} = \langle \mathcal{P}, \mathcal{I}, \mathcal{A}, \preceq \rangle$ is determined by a nonempty set $\mathcal{P}$ of basic propositions; by a nonempty set $\mathcal{A}$ of agent indices; by a function $\mathcal{I}$ on $\mathcal{A}$, where $\mathcal{I}_a$ is a nonempty set of subagents (the subagents of a); and by a function $\preceq$ which for each $a \in \mathcal{A}$ provides a reflexive, transitive ordering on $\mathcal{A}$.

[29] This section corresponds to parts of [Thomason, 1998].

I do not assume that if $a \neq b$, then $\mathcal{I}_a$ and $\mathcal{I}_b$ are disjoint; in fact, we often want to consider agents with the same general epistemic organization, and in this case, $\mathcal{I}_a = \mathcal{I}_b$ for all a and b.

Both the pure intra-agent logic and the subagentless multi-agent epistemic logic are special cases of inter-agent modal logic. We obtain the familiar multi-agent case by letting $\mathcal{I}_a = \{i_a\}$ for all $a \in \mathcal{A}$. We obtain the pure intra-agent case by letting $\mathcal{A} = \{a\}$.

Definition 6.5. *Inter-Agent Modal Systems with Mutual Belief.*
An inter-agent propositional language $\mathcal{L} = \langle \mathcal{P}, \mathcal{I}, \mathcal{A}, \preceq, \textsc{mut} \rangle$ with mutual belief attitudes is a an agent-homogeneous inter-agent propositional language with a modal operator $[\,\textsc{mut}, i\,]$ for each $i \in \mathcal{I}$. The satisfaction condition for $[\,\textsc{mut}, j\,]$ is as follows:

$M \models_{a,i,w} [\,\textsc{mut}, j\,]A$ iff $i \preceq j$ and $M \models_{a,i,w'} A$ for all w' such that $wR_{c,j}w'$, where R_c is the transitive closure of the set of relations $\{R_{b,j} : a \in \mathcal{A}\}$.

The resulting logic contains standard multi-agent modal logics for reasoning about mutual belief, such as the system KD45$_n^C$ of [Fagin *et al.*, 1995].

7. An Example of Intra-Agent Modality: Keeping Track of Public and Private Beliefs

Note: This section corresponds to part of [Thomason, 1998].

In general, we find it useful not only to believe many things, but to keep track of which of these beliefs are public and which are not. If it is public knowledge where my car is, I can tell you I'll meet you in fifteen minutes at my car. If it is not, I will have to tell you where my car is. And I will have to do this in public terms.

Example 1.

The simplest example I can think of invokes only three subagents: PUB, NPUB and the agent MIN combining beliefs from both of these sources. Then the reflexive relation $\preceq$ on subagents has five elements:

$$\preceq = \{\langle \textsc{min}, \textsc{min} \rangle, \langle \textsc{min}, \textsc{pub} \rangle, \langle \textsc{min}, \textsc{npub} \rangle, \langle \textsc{pub}, \textsc{pub} \rangle, \langle \textsc{npub}, \textsc{npub} \rangle\}.$$

Suppose that our language has just two basic propositions,

(p_1) MONDAY(TODAY)
(p_2) MY-BIRTHDAY(TODAY)

There are three worlds:

w_1: p_1 is true and p_2 is true.
w_2: p_1 is true and p_2 is false.
w_3: p_1 is false and p_2 is true.

Accessibility is defined as follows.

$R_{\textsc{npub}}(w, w')$ iff $w, w' \in \{w_1, w_3\}$.
$R_{\textsc{pub}}(w, w')$ iff $w, w' \in \{w_1, w_2\}$.

Supposing that w_1 is the actual world and that MIN represents the compiled beliefs of the agent, satisfaction at the "viewpoint" $\langle \textsc{min}, w \rangle$ will represent what holds in the actual world for the agent. The following formulas hold here.

1. $p_1 \wedge p_2$
2. $[\,\text{NPUB}\,]p_2$, $\neg[\,\text{NPUB}\,]p_1$
3. $[\,p, \text{PUB}\,]_1$, $\neg[\,\text{NPUB}\,]p_2$
4. $[\,\text{MIN}\,][p_1 \wedge p_2]$, $[\,\text{MIN}\,][\,\text{NPUB}\,]p_2$, $[\,\text{MIN}\,]\neg[\,\text{NPUB}\,]p_1$, $[\,\text{MIN}\,][\,p, \text{PUB}\,]_1$, $[\,\text{MIN}\,]\neg[\,\text{NPUB}\,]p_2$

I hope that this simple example will make clear the usefulness of the formalism in representing how agents might keep track of public and private information. Ordinarily, I expect anyone that I meet to share my beliefs about what day of the week it is. But there are only a few people whom I would expect to be aware of my birthday. The formalism enables us to represent these distinctions. For instance, the formulas in Line 4 say that (i) I believe that today is Monday and my birthday, (ii) I believe that it is a private belief that today is my birthday, (iii) I believe that it is not a private belief that today is Monday, (iv) I believe that it is a public belief that today is Monday, and (v) I believe that it is not a private belief that today is Monday. And these distinctions are represented in a way that uses the familiar modal apparatus for representing the epistemic attitudes.

Note that even if we are careful to control the information that goes into the NPUB module by *not* putting axioms into it that go into the PUB module, it will contain at least some public information. (For instance, any tautology will be known by any module.) So the fact that $[\,\text{NPUB}\,]A$ holds does not in itself prevent A from expressing some piece of public information.

8. Achieving mutuality through nonmonotonic reasoning[30]

8.1. Simplifications

We will assume that there are only two agents, and that each agent has only one minimal subagent i_0 (which is also public). So there are only two agent modalities: $[\,a, i_0\,] = [\,a, \text{PUB}\,]$ and $[\,b, i_0\,] = [\,b, \text{PUB}\,]$. This simplification will enable us to work without the apparatus of quantificational logic. They also enable some simplifications in the model theory. Where there is only one subagent per agent, satisfaction in a model does not depend on a choice of any particular subagent. So we can revert to the familiar case in which satisfaction in a model depends only on the choice of world.

[30] This section corresponds to parts of [Thomason, 2000].

8.2. Introducing nonmonotonicity

There are a number of formalizations of nonmonotonic logic. I will use a circumscriptive approach[31] here. The most general formulations of circumscription present the logical ideas model theoretically, using a *preferred models* approach. The ideas are similar to those used in the semantics of conditionals, but in this case one works with models rather than with possible worlds.

Preferred models approaches postulate a relation $\preceq$ over models. Nonmontonic logical consequence is then defined as follows:

(1) M is a maximally preferred model of Γ if and only if M is a model of Γ and for all models M' of Γ such that $M' \preceq M$, $M' = M$.

(2) A set Γ of formulas nonmonotinically implies a formula A if and only if A is true in all the maximally preferred models of Γ.

In a circumscriptive logic, $\preceq$ is defined by simultaneously minimizing certain abnormalities, while the extensions of some other terms are held constant, and those of still other terms are allowed to vary. (What is held constant and what is allowed to vary has to be decided in terms of the particular application.)

Usually, these abnormalities are the extensions of first-order abnormality predicates, which are used in axiomatizing the nonmonotonic theory. Typically, a generalization of the form "*ceteris paribus*, P's are Q's" would be formalized along the following lines in a circumscriptive theory.

$$\forall x[[P(x) \wedge \neg Ab(x)] \rightarrow Q(x)].$$

Here, Ab is a first-order abnormality. In the case of reasoning about attitudes, we will be concerned with abnormalities that take propositional arguments.

Very little work has been done on modal circumscriptive logics. But nothing intrinsic to circumscription prevents us from minimizing abnormalities concerning modalities. As we will see, we will be interested in minimizing the difference between an agent **a**'s public beliefs and **a**'s view of **b**'s public beliefs. More precisely, we will want to minimize these differences while allowing **a**'s (public) beliefs about **b**'s (public) beliefs to vary while **b**'s actual beliefs and **a**'s beliefs about other matters are held constant. Hopefully, in cases where nothing interferes, the maximally preferred models should be ones in which a believes (publically) that **a**'s and **b**'s (public) beliefs are mutual.

There is a technical problem here which deserves careful attention, and which is discussed in detail in [Thomason, 2000], but which I will not stress here. We wish to allow **a**'s beliefs about **b**'s beliefs to vary while **b**'s actual beliefs and **a**'s beliefs about other matters are held constant. But a semantics that uses relations R_a and R_b over possible worlds to do this doesn't separate the two kinds of beliefs. The problem is that, for instance, nothing prevents both $w_0 R_a w$ and $w_0 R_b w$. But then we couldn't change the worlds R_b related to w in order to minimize **a**'s beliefs about **b**'s beliefs without changing **b**'s actual beliefs. To deal with this, I show that every model is equivalent to a separated model in which this sort of problem doesn't arise, and work with separated models.

[31] See [McCarthy, 1980, Lifschitz, 1994].

The following definition provides more detail concerning the formalization of the preference relation for the case in which we are concerned merely with the beliefs of agent **a**.

Definition 8.6. *$M_1 \cong_a M_2$, $M_1 \leq_a M_2$, a-minimality for T , $\|\!\!\sim_{\mathrm{a}}$.*

Let $\boldsymbol{Ab} = \{Ab^a_1\}$, $\mathcal{G}' = \{a\}$, $\mathcal{G} = \{a, b\}$, and $i = i_0$. Then:

(1) $M_1 \cong_a M_2$ iff $M_1 \cong_{\boldsymbol{Ab},\mathcal{G},\mathcal{G}',i} M_2$;
(2) $\mathcal{M}_1 \leq_a \mathcal{M}_2$ iff $\mathcal{M}_1 \leq_{\boldsymbol{Ab}} \mathcal{M}_2$;
(3) $\mathcal{M}$ is a-minimal for T iff $\mathcal{M}$ is $Ab, \mathcal{G}, \mathcal{G}', i$-minimal for T;
(4) $T \|\!\!\sim_{\mathrm{a}} A$ iff $T \|\!\!\sim_{\boldsymbol{Ab},\mathcal{G},\mathcal{G}',\mathrm{i}} A$.

8.3. Epistemic transfer axiom schemata

Since each agent has only one subagent, we can simplify the notation for modalities: for instance, we let $[\,a\,] = [\,a, \textsc{pub}\,]$.

The following epistemic transfer axiom schemata provide an incremental approach to the reasoning that underlies mutuality.

$$(\mathrm{Tr}_a) \quad [[\,a\,]A \wedge \neg Ab^a_1(A)] \rightarrow [\,a\,][\,b\,]A$$

$$(\mathrm{Tr}_b) \quad [[\,b\,]A \wedge \neg Ab^b_1(A)] \rightarrow [\,b\,][\,a\,]A$$

According to Axiom Schema (Tr_a), ***a*** normally believes that ***b*** believes whatever ***a*** believes; Axiom Schema (Tr_b) says the corresponding thing about **b**. We can show that if the extension of $Ab^a_1(A)$ is empty for all A, Axiom Schema (Tr_a) implies that whatever **a** publically believes, **a** also believes to be mutual. (And similarly, of course, for Axiom Schema (Tr_b) and **b**.) I will use 'Tr_a' and 'Tr_b' to denote not only the schemata given above, but the corresponding sets of formulas.

8.4. A result about mutual belief

Let Normal-1 $= \{\neg Ab^a_1(A) :\ A$ a formula$\}$. Consider the following rule concerning mutual belief.

(8.1) From $\mathrm{Tr}_a \cup \{$Normal-1$\}$ infer $[\,a\,]A \rightarrow [\,a\,][\,\textsc{mut}\,]A$

Lemma 8.1. The rule (8.1) is valid.

This lemma is proved in [Thomason, 2000]. We can show that under fairly general conditions (but not nearly as general as I would like), Lemma 8.1 and the transfer axioms ensure that, under general conditions, agents will by default believe that their own beliefs are mutual. This does not, of course, imply that the beliefs of agents will in fact, even by default, be mutual. And this is not something we should expect to prove. Suppose, for example, that **a** and **b**, while standing side by side, read a poster saying that a lecture will be given at 9, believing what they read. Also suppose that the transfer axiom schemata hold for these agents. But **a** is a morning person, and believes that the lecture will be given at 9am, while **b** is a night person, and believes the lecture will be given at 9pm. Then (assuming that neither agent is aware of a relevant abnormality, and in particular has not noticed the ambiguity in the poster) each agent will believe that their beliefs are mutual, but this belief will be incorrect. The agents will be *uncoordinated*, in the sense that their mutually modeling public modules will in fact differ.

The transfer axiom schemata allow cases of this kind to occur without any concommitant abnormalities. That is, although epistemic transfer creates defaults about agent's beliefs about what they each believe, it creates no defaults about the coordination of agents. To put it another way, the transfer axiom schemata only apply to what agents believe about one another. They do not apply to what a third party should believe about the agents' beliefs about the world, so they will not enable us to infer epistemic coordination of a group by default. There are, of course, circumstances under which a third party could have reason to suppose that a group of agents is coordinated, but I do not attempt to formalize these.

The following is a rough statement of a theorem that is stated and proved in [Thomason, 2000].

Theorem 8.1. Let T be a theory that contains no statements concerning **a**'s beliefs about **b**'s beliefs, or about abnormalities, and let A be a formula also meeting these conditions. Then $T \mathrel{\|\!\!\mid\!\!\!\sim_{\mathrm{a}}} \texttt{[}\,a\,\texttt{]}A \rightarrow \texttt{[}\,a\,\texttt{]}\texttt{[}\,\textsc{mut}\,\texttt{]}A$, for all formulas A.

9. An Example

Belief transfer is fallible, and is recognized as such in everyday cases of reasoning about belief. So it is important to provide a means of formalizing the circumstances under which the defeasible leap to mutuality will be blocked.

The example I'll develop in this section resembles the one at the beginning of [Clark and Marshall, 1981] which shows that the iteration complexity for agent knowlege can reach fairly high finite levels, and that these levels do not support reference presuppositions. That example, though, deals simply with agent beliefs. As I explained in Section 3.3.2, I believe that agent beliefs are the wrong attitudes to use when mutuality is at stake. Instead, we need "public" attitudes that are invoked specifically to model other agents.

As usual, the following example involves two agents, a and b. We distinguish between their private beliefs and the beliefs that they expect to be public in a conversation they are having along a potentially faulty communication channel. Each agent has two subagents, ROOT and PUB. The former represents the sum of the agent's beliefs and the latter represents the belief module that is devoted to tracking the conversation. We have ROOT $\preceq$ PUB, but PUB $\not\preceq$ ROOT.

The following rudimentary theory of email communication consists of three parts: (A) protocols for updating the contents of $\texttt{[}\,a, \textsc{pub}\,\texttt{]}$, (B) a theory of exceptions to the protocols in (1), and (C) the transfer axioms (Tr_a) and (Tr_b). To keep things simple, the formalization ignores temporal considerations.

(A) **Protocols for updating** $\texttt{[}\,a, \textsc{pub}\,\texttt{]}$

(A.1) $\forall m \forall p[[\mathit{Send}(a,b,m) \wedge \mathit{Incontents}(m,p)] \rightarrow \texttt{[}\,a, \textsc{pub}\,\texttt{]}p]$

If **a** sends a message to **b** that says p then **a** adds < > to $\texttt{[}\,a, \textsc{pub}\,\texttt{]}$.

(A.2) $\forall m_1 \forall m_2 \forall p[[\mathit{Send}(a, b, m_1) \wedge \mathit{Incontents}(m_1, p) \wedge \mathit{Read}(a, m_2)$
$\qquad \wedge\, \mathit{sender}(m_1) = b \wedge \mathit{Ack}(m_2, m_1)] \rightarrow [a, \mathrm{PUB}][b, \mathrm{PUB}]p]$

If **a** sends a message to **b** that says p and reads an acknowledgement of that message from b then **a** adds $[b, \mathrm{PUB}]p$ to $[a, \mathrm{PUB}]$.

(B) **The abnormality theory.**

(B.1) $\forall m_1 \forall m_2 \forall p[\mathit{Send}(a, b, m_1) \wedge \mathit{Incontents}(m_1, p) \wedge \mathit{Read}(a, m_2)$
$\qquad \wedge\, \mathit{Not\text{-}Delivered}(m_2, m_1)] \rightarrow \mathit{Ab}_1^a(p)]$

If **a** sends a message to **b** saying that p and reads a message saying the message was not delivered then the conditions for inferring $[b, \mathrm{PUB}]p$ are blocked.

(B.2) $\forall m_1 \forall m_2 \forall m_3 \forall p[[\mathit{Send}(a, b, m_1) \wedge \mathit{Incontents}(m_1, p) \wedge \mathit{Read}(b, m_2)$
$\qquad \wedge\, \mathit{Ack}(m_2, m_1) \wedge [a, \mathrm{ROOT}][\mathit{Read}(b, m_3)$
$\qquad \wedge\, \mathit{Not\text{-}Delivered}(m_3, m_2)]] \rightarrow \mathit{Ab}_1^a([a, \mathrm{PUB}]p)$

If **a** sends a message to **b** that says p and reads an acknowledgement of that message from **b** and believes that **b** receives a message saying that the acknowledgement was not delivered then the conditions for inferring that $[b, \mathrm{PUB}][a, \mathrm{PUB}]p$ are blocked.

(C) **The transfer axiom scheme**

(Tr_a) $[[[a, \mathrm{PUB}]A \wedge \neg \mathit{Ab}_1^a(A)] \rightarrow [a, \mathrm{PUB}][b, \mathrm{PUB}]A]$

I will present three cases, of increasing complexity. In the simplest case, mutuality is inferred.

Case 1. The story: Ann and Bob correspond regularly and normally by email. Ann sends Bob the following message, M_1. Nothing unusual happens.

```
To: Bob <robert@xyz.org>
From: Ann <ann@abc.org>
Subject: Movies at the Roxie

Bob,
Monkey Business is showing tonight at the Roxie.
Ann
```

The reasoning: Since her communications with Bob are normally successful, Ann assumes that this one is successful, and in fact it has much the same status for her that face-to-face conversation does. Ann maintains a subagent to keep track of beliefs that are *prima facie* shared with Bob. On sending M_1, she adds the contents of the message to the beliefs of this subagent, i.e., MB is added to $[a, \mathrm{PUB}]$. Although the matter is more complicated than I would like it to be, I believe it can be shown that in this case $[a, \mathrm{PUB}][\mathrm{MUT}]\mathit{MB}$ will be a circumscriptive consequence.

The formalization:

Initial Conditions: $\mathit{Send}(a, b, M_1, e_1)$, $\mathit{Incontents}(M_1, \mathit{MB})$

Monotonic consequence: $[\,a, \textsc{pub}\,]\mathit{MB}$

Circumscriptive consequence: $[\,a, \textsc{pub}\,][\,\textsc{mut}\,]\mathit{MB}$.

Case 2. The story: Ann sends the following message, M_2, to Bob.

```
To: Bob <bob@xyz.org>
From: Ann <ann@abc.org>
Subject: Movies at the Roxie

Bob,
Monkey Business is showing tonight at the Roxie.
Ann
```

Immediately afterwards, she receives the following message, M_3. She says to herself "Oops, I misaddressed the message."

```
To: ann@abc.org
From: mailer-daemon@xyz.org
Subject: undeliverable mail

The following errors occurred when trying to deliver the
attached mail:

bob: User unknown
```

The reasoning: As in Case 1, Ann adds the contents of M_1 to the beliefs of the subagent representing *prima facie* beliefs shared with Bob. However, the receipt of the mailer daemon's message precipitates an anomaly, which in turn blocks any ascription of this belief to Bob.

The formalization:

Initial Conditions:

$\mathit{Send}(a, b, M_2, e_1)$, $\mathit{Incontents}(M_2, \mathit{MB})$
$\mathit{Read}(a, M_3)$, $\mathit{Not\text{-}Delivered}(M_3, M_2)$

Consequences: $[\,a, \textsc{pub}\,]\mathit{MB}$ is a consequence, but $[\,a, \textsc{pub}\,][\,b, \textsc{pub}\,]\mathit{MB}$ is not.

Case 3. The story: Ann has just returned from a vacation. Forgetting to turn off her vacation daemon, she sends the misaddressed message M_2 to Bob. She receives error message M_3 from the mailer daemon. Shortly after that, she receives the following message, M_4, from Bob.

```
To: ann@abc.org
From: Bob <robert@xyz.org>
Subject: re: Movies at the Roxie

Ann,
We just rigged the mailer here to send me blind
copies of messages to bob@xyz.org, so actually I got your
message about Monkey Business.
Bob
```

She realizes that Bob has received an automatic reply to M_4 from her vacation daemon saying that she is on vacation, but will answer the message as soon as she gets back.

The reasoning: As in Cases 1 and 2, Ann adds MB to the beliefs of the subagent representing *prima facie* beliefs shared with Bob. But the receipt of the mailer daemon's message precipitates an anomaly, which in turn blocks any default ascription of this belief to Bob. Bob's acknowledgement overrides this anomaly, allowing Ann's public subagent to conclude that Bob's public subagent believes MB. But when she learns that Bob received a message indicating that she didn't receive his acknowledgement, an anomaly is generated which blocks the inference in Ann's public subagent that Bob's public subagent believes that Ann's public subagent MB.

I omit the details of the formalization of this example. If it is formalized properly, axioms will prevent $[\textsc{mut}]MB$ from being added to $[a, \textsc{pub}]$. In fact, $[a, \textsc{pub}]$ should contain $[b, \textsc{pub}]MB$ but not $[b, \textsc{pub}][a, \textsc{pub}]MB$.

10. Conclusion

Previous attempts to model reasoning about knowledge have not provided plausible formalizations of the reasoning that underlies mutuality in cases that seem to require it, or provided logical materials for formalizing cases where mutuality is blocked. Unless I have missed something, the literature contains no flexible formal reasoning mechanisms for obtaining mutuality.

Many authors have suggested that mutuality somehow arises spontaneously out of certain shared situations. This suggestion is flawed, since shared situations do not in general lead to mutuality—for instance, I will not treat information that I obtain from a situation I share with you as mutual if I observe that you do not observe me sharing the situation. If we believe that mutuality is required for some purposes, then we have to produce a reasoning mechanism that allows agents to obtain it from information that we can plausibly expect agents to have, and that also allows us to block the reasoning in cases where mutuality should not be forthcoming.

The only way to demonstrate the viability of a theory of these mechanisms is to demonstrate their utility in formalizing a wide variety of fairly complex cases. I have not done that here. But I hope that at least I have made a plausible case for the promise of the approach that is developed in this paper.

I believe that the theory also offers hints about the cognitive foundations of mutuality and social attitudes that are more detailed and promising than any other models I am aware of. But what I have presented is a logical theory, not a cognitive model.

We also need implementations of this sort of reasoning, if computers are to function effectively in cooperative groups that contain humans. The theory I have presented is not, of course, an implementation. In particular, it does not show how the reasoning can be efficiently implemented in special cases.

Bibliography

[Aumann, 1976] Robert J. Aumann. Agreeing to disagree. *Annals of Statistics*, 4(6):1236–1239, 1976.

[Bar-Tal and Kruglanski, 1988a] Daniel Bar-Tal and Arie W. Kruglanski, editors. *The Social Psychology of Knowledge*. Editions de la Maison des Sciences de l'Homme, Cambridge, England, 1988.

[Bar-Tal and Kruglanski, 1988b] Daniel Bar-Tal and Arie W. Kruglanski. The social psychology of knowledge: Its scope and meaning. In Daniel Bar-Tal and Arie W. Kruglanski, editors, *The Social Psychology of Knowledge*, pages 1–14. Editions de la Maison des Sciences de l'Homme, Cambridge, England, 1988.

[Barwise, 1988] K. Jon Barwise. Three views of common knowledge. In Moshe Y. Vardi, editor, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, pages 365–379, Los Altos, California, 1988. Morgan Kaufmann.

[Buvač and Mason, 1993] Saša Buvač and Ian Mason. Propositional logic of context. In Richard Fikes and Wendy Lehnert, editors, *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 412–419, Menlo Park, California, 1993. American Association for Artificial Intelligence, AAAI Press.

[Clark and Marshall, 1981] Herbert H. Clark and Catherine R. Marshall. Definite reference and mutual knowledge. In Arivind Joshi, Bonnie Webber, and Ivan Sag, editors, *Linguistics Structure and Discourse Setting*, pages 10–63. Cambridge University Press, Cambridge, England, 1981.

[Clark and Schober, 1989] Herbert H. Clark and Michael Schober. Understanding by addressees and overhearers. *Cognitive Psychology*, 24:259–294, 1989. Republished in [Clark, 1992b].

[Clark, 1992a] Herbert Clark. *Arenas of Language Use.* University of Chicago Press, Chicago, 1992.

[Clark, 1992b] Herbert Clark. *Arenas of Language Use.* University of Chicago Press, Chicago, 1992.

[Cole, 1991] M. Cole. Conclusion. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 398–417. American Psychological Association, Washington, D.C., 1991.

[Davies and Stone, 1995a] Martin Davies and Tony Stone, editors. *Folk Psychology: The Theory of Mind Debate.* Blackwell, Oxford, 1995.

[Davies and Stone, 1995b] Martin Davies and Tony Stone, editors. *Mental Simulation: Evaluations and Applications.* Blackwell, Oxford, 1995.

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

[Fagin, 1994] Ronald Fagin, editor. *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the Fifth Conference (TARK 1994).* Morgan Kaufmann, San Francisco, California, 1994.

[Geanakoplos, 1990] John Geanakoplos. Common knowledge in economics. In Rohit Parikh, editor, *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the Third Conference (TARK 1990)*, pages 139–140, Los Altos, California, 1990. Morgan Kaufmann. This is an abstract for a tutorial session.

[Geanakopolos, 1992] John Geanakopolos. Common knowledge. In Yoram Moses, editor, *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the Fourth Conference (TARK 1992)*, pages 254–315. Morgan Kaufmann, San Francisco, 1992.

[Gilbert, 1989] Margaret Gilbert. *On Social Facts.* Routledge, London, 1989.

[Guha, 1991] Ramanathan V. Guha. Contexts: a formalization and some applications. Technical Report STAN-CS-91-1399, Stanford Computer Science Department, Stanford, California, 1991.

[Halpern, 1986] Joseph Y. Halpern, editor. *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the First Conference (TARK 1986).* Morgan Kaufmann Publishers, Inc., Los Altos, California, 1986.

[Hintikka, 1962] Jaakko Hintikka. *Knowledge and Belief.* Cornell University Press, Ithaca, New York, 1962.

[Kripke, 1965] Saul Kripke. A semantical analysis of modal logic ii: Non-normal propositional calculi. In Leon Henkin and Alfred Tarski, editors, *The Theory of Models*, pages 206–220. North-Holland, Amsterdam, 1965.

[Lemmon, 1957] E.J. Lemmon. New foundations for Lewis modal systems. *Journal of Symbolic Logic*, 22(2):176–186, 1957.

[Lewis, 1969] David K. Lewis. *Convention: A Philosophical Study.* Harvard University Press, Cambridge, Massachusetts, 1969.

[Lewis, 1979] David K. Lewis. Scorekeeping in a language game. *Journal of Philosophical Logic*, 8(3):339–359, 1979.

[Lifschitz, 1994] Vladimir Lifschitz. Circumscription. In Dov Gabbay, C. J. Hogger, and J. A. Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming, Volume 3: Nonmonotonic Reasoning and Uncertain Reasoning*, pages 298–352. Oxford University Press, 1994.

[Marks and Miller, 1987] Gary Marks and Norman Miller. Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, 102:72–90, 1987.

[McCarthy, 1977] John McCarthy. Epistemological problems of artificial intelligence. In *Proceedings of the Fifth International Joint Conference on AI*, pages 1038–1044, Pittsburgh, PA, 1977. Department of Computer Science, Carnegie Mellon University.

[McCarthy, 1980] John McCarthy. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence*, 13(1–2):27–39, 1980.

[Milgrom and Stokey, 1982] P. Milgrom and N. Stokey. Information, trade, and common knowledge. *Journal of Economic Theory*, 26:17–27, 1982.

[Mortensen, 1996] C. David Mortensen. *Miscommunication.* Sage Publications, Thousand Oaks, California, 1996.

[Moses, 1992] Yoram Moses, editor. *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the Fourth Conference (TARK 1992).* Morgan Kaufmann, San Francisco, California, 1992.

[Parikh, 1990a] Rohit Parikh. Recent issues in reasoning about knowledge. In Rohit Parikh, editor, *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the Third Conference (TARK 1990)*, pages 3–10, Los Altos, California, 1990. Morgan Kaufmann.

[Parikh, 1990b] Rohit Parikh, editor. *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the Third Conference (TARK 1990).* Morgan Kaufmann, Los Altos, California, 1990.

[Resnick *et al.*, 1991] Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors. *Perspectives on Socially Shared Cognition.* American Psychological Association, Washington, D.C., 1991.

[Schiffer, 1972] Stephen Schiffer. *Meaning.* Oxford University Press, Oxford, 1972.

[Shoham, 1996] Yoav Shoham, editor. *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the Sixth Conference (TARK 1996).* Morgan Kaufmann, San Francisco, California, 1996.

[Stalnaker, 1975] Robert C. Stalnaker. Pragmatic presuppositions. In Miltin K. Munitz and Peter Unger, editors, *Semantics and Philosophy.* Academic Press, New York, 1975.

[Thielscher, 1996] Michael Thielscher. Causality and the qualification problem. In Luigia Carlucci Aiello, Jon Doyle, and Stuart Shapiro, editors, *KR'96: Principles of Knowledge Representation and Reasoning*, pages 51–62. Morgan Kaufmann, San Francisco, California, 1996.

[Thomason, 1990] Richmond Thomason. Accommodation, meaning, and implicature: Interdisciplinary foundations for pragmatics. In Philip R. Cohen, Jerry Morgan, and Martha Pollack, editors, *Intentions in Communication*, pages 326–363. MIT Press, Cambridge, Massachusetts, 1990.

[Thomason, 1998] Richmond H. Thomason. Intra-agent modality and nonmonotonic epistemic logic. In Itzhak Gilboa, editor, *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the Seventh Conference (TARK 1998)*, pages 57–69, San Francisco, California, 1998. Morgan Kaufmann.

[Thomason, 2000] Richmond H. Thomason. Modeling the beliefs of other agents. In Jack Minker, editor, *Logic-Based Artificial Intelligence*, pages 375–473. Kluwer Academic Publishers, Dordrecht, 2000.

[Tuomela, 1995] Raimo Tuomela. *The Importance of Us.* Stanford University Press, 1995.

[Turing, 1936] Alan M. Turing. On computable numbers with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society Series 2*, 42:230–265, 1936.

[Vardi, 1988] Moshe Y. Vardi, editor. *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the Second Conference (TARK 1988).* Morgan Kaufmann, San Francisco, California, 1988.