Modeling the Beliefs of Other Agents

Richmond H. Thomason
AI Laboratory
University of Michigan
Ann Arbor, MI 48109–2210
U.S.A.

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

Version of: November 4, 2000

1. Introduction and background

The larger project of which this paper is a part is an attempt to develop a logical framework for *agent modeling*, the business of formulating and maintaining representations of the attitudes of other agents. This is something that normal, adult humans do very well. Take reasoning about beliefs. Under a wide variety of circumstances, we are able to reach rather detailed hypotheses about what the people with whom we interact believe, and to modify these hypotheses in the light of new evidence. But for some reason, it is especially difficult to obtain reflective insight into how we do this reasoning, to break it down into steps and to formalize it.

This paper is devoted to reasoning about belief and belief-like attitudes, and especially to the problem of how agents can achieve mutuality by this reasoning.[1]

1.1. The need for mutuality

I will not devote much space in this paper to motivating the importance in various human endeavors of mutuality (of knowledge, or belief, or of other similar attitudes), or to arguing for the plausibility of the hypothesis that groups are often capable of achieving something like mutuality. The need for mutuality seems to turn up whenever conscious, human-level interactions are analyzed at a sufficiently deep level. It shows up in philosophical accounts of convention Lewis [14], in accounts of conversation Clark & Marshall [5], and in economic analyses of interactions Aumann [1]. Maybe it will suffice to give just one example of the importance of mutuality in thinking about such phenomena: you can't model a game like stud poker without accounting for the difference between the cards that are dealt face up and the ones that are dealt face down. The difference is simply that the values of the face-up cards are mutually known, and the values of the face-down cards are not.

1.2. An opportunity for logic-based AI

I believe that agent modeling presents special opportunities for logic-based AI. That is, I believe that the techniques that have been developed in logic-based AI are well suited for formalizing much of the reasoning that is involved in agent modeling, and that the probabilisitic techniques which work so well in modeling many natural phenomena are not likely to work well in many cases of agent modeling. The main purpose of this paper is to establish at least the positive point—I will try to show that a combination of modal logic and nonmonotonic logic provides a suitable foundation for agent modeling.

[1] Readers who are not familiar with the term 'mutuality' should consult Section 2.1, below. There are terminological differences in the literature—some authors use 'common', as in 'common ground' and 'common knowledge', others use 'mutual', as in 'mutual belief'. I prefer the latter term, because it nominalizes more happily: 'mutuality' sounds better than 'commonality'.

2. The problem of achieving mutuality

2.1. What is mutuality?

In this paper p, q, and r are used for propositions, and A, B and C for syntactic objects, usually sentences of a natural language or formulas of a formalized language. Propositions are to be understood as objects of propositional attitudes. The theory presented in this paper subscribes to a formalization of propositions that identifies them with sets of possible worlds. In the following, where a sentence or formula A is mentioned in connection with a proposition p, it is assumed that A expresses p.

Take a group G of agents. Suppose that this group has attitudes—say beliefs—and that among these are beliefs about the beliefs of other agents. In that case, we can talk about mutual belief for the group. A proposition p is *mutually believed* by the group in case every member of the group believes p, and believes that every member believes p, and believes that every member believes that every member believes p, and so forth. So, for instance, if p is mutually believed by a group G including members $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$, then $\boldsymbol{a}$ believes that $\boldsymbol{c}$ believes that $\boldsymbol{b}$ believes that $\boldsymbol{a}$ believes that p. Similarly, a sentence or formula is mutually believed if it expresses a proposition that is mutually believed.

Using notation from modal logic, I mark an attitude relating an agent $\boldsymbol{a}$ to a proposition p by $[a]A$. From here on, except in a few places where it is convenient to state things generally, I will confine myself to groups containing just two agents, $\boldsymbol{a}$ and $\boldsymbol{b}$.

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, ϕ is the empty string.

We make the idea of an iterated attitude, and its connection to mutuality, precise in the following definitions.

Definition 2.1. *Iteration depth.*

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

Definition 2.2. *Iteration complexity.*
The $[a]$-iteration complexity of A in a model or example is the smallest ordinal $\eta \leq \omega$ such that A does not iterate to depth $\eta + 1$ in the model or example.
$[b]$ iteration complexity is defined analogously to $[a]$ iteration complexity.

Mutuality is characterized globally for groups; for groups of more than one agent, it can't be defined in terms of the attitudes of any single agent. But reasoning is performed by individuals and is agent-bound. In this paper, where the focus is on reasoning, it is natural to concentrate on the iteration complexity for the attitudes of a single, fixed agent. I take this approach in the discussion below, and in the subsequent formal developments. Fortunately, mutuality can be characterized in terms of the attitudes of the separate agents in a group: p will be mutually believed if and only if the iteration complexity of A is infinite for every member of the group.

In an example where $[a]A$ is false, the $[a]$ iteration complexity of A is 0. If $[a]A$ is true but $[a][b]A$ is false, the $[a]$ iteration complexity of A is 1. If $[a]A$ and $[a][b]A$ are true but $[a][b][a]A$ is false, the $[a]$ iteration complexity of A is 2. If the $[a]$ and the $[b]$ iteration complexity of A are both ω, A is mutual for the group $\{\boldsymbol{a}, \boldsymbol{b}\}$ and the attitude $[\,]$. If the $\boldsymbol{a}$-*believes* and the $\boldsymbol{b}$-*believes* iteration complexity of A is greater than 0, but finite, we have a case in which both $\boldsymbol{a}$ and $\boldsymbol{b}$ 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 $\boldsymbol{a}$-*believes* and the $\boldsymbol{b}$-*believes* iteration complexities of A 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.

2.2. Immediacy theories

I take it for granted that we want a formalized theory of the reasoning processes that yield mutual attitudes, and that a multimodal logic is the appropriate vehicle for formalizing the relevant agent attitudes.

The most detailed studies with which I am familiar of the reasoning leading to mutual attitudes[2] all suggest that an immediacy-based approach can account for the reasoning. The three accounts are similar, partly because the two later ones are influenced by Lewis. None of them is fully formalized; they all suggest that mutuality is somehow precipitated by the reflexive character of certain shared situations. There are differences in the three versions, which would probably lead to different formalizations; Barwise [2] suggests different formalizations of Lewis [14] and Clark & Marshall [5]. However, I don't believe that these differences are relevant to the points I wish to make.

I believe that intuitions about immediacy fail to provide an adequate formalization of the relevant reasoning. I'll use a simplified version of the immediacy theory to explain the point. Although it doesn't correspond exactly to any earlier presentation of the idea and won't do full justice to 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 sentence or formula A *guarantees B-mutuality* for agent $\boldsymbol{a}$ and group $\{\boldsymbol{a}, \boldsymbol{b}\}$ in an example or model if

$$\textbf{MutAx} \qquad [a][A \rightarrow [\textsc{mut}]B]$$

holds in that example or model. A proposition p guarantees B-mutuality for agent $\boldsymbol{a}$ and group $\{\boldsymbol{a}, \boldsymbol{b}\}$ in an example or model if A guarantees B-mutuality, where, as usual, A expresses p.

The idea behind immediacy is that agents apparently often find themselves in circumstances where something like MutAx can be used to secure mutuality. The following sorts

[2] By David Lewis [14], Stephen Schiffer [21], and Herbert Clark and Catherine Marshall [5].

of examples are used to illustrate the point. In Example 1, (1a) represents A in MutAx and (1b) represents B.[3] In Example 2, (2a) represents A in MutAx and (2b) represents B.[4]

(1a) $\boldsymbol{a}$ and $\boldsymbol{b}$ 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) $\boldsymbol{b}$ says to $\boldsymbol{a}$, 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) $\boldsymbol{b}$ will be at l at 9am on $d + 1$.

In Example 1, what guarantees mutuality for (1b) is the proposition that $\boldsymbol{a}$ and $\boldsymbol{b}$ are face-to-face across a small table, and are both looking at the candle—or maybe a qualification of this proposition.[5] In Example 2, what guarantees mutuality for (2b) is the proposition that, in a face-to-face conversation, $\boldsymbol{a}$ has just said to $\boldsymbol{b}$ "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 MutAx *guarantees* this inference. I'll divide my explanation of why I think that immediacy theories provide an inadequate account of the reasoning into three topics: *incrementality*, *monolithicity*, and *fallibility*.

2.2.1. Incrementality issues

Finite iteration complexity can arise in naturally occurring situations. A classic series of examples of increasing finite complexity is presented in Clark & Marshall [5]; another is developed in Section 9, below, where for a certain A, $[a][b]A$ holds but $[a][b][a]A$ 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 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.

2.2.2. Monolithicity issues

MutAx implies $[a]A \rightarrow [a][b]A$ and $[a]A \rightarrow [a][b][a]A$. 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; the best she can do for Bob is 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 says, "Now,

[3] The example is from Lewis [14, pp. 52–57].

[4] An example from Schiffer [21, pp. 31–36].

[5] Schiffer produces a fairly elaborate qualification, Schiffer [21, p. 35].

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. Nevertheless, the interview takes place without any presuppositional anomalies, without any of the abnormalities that accompany failures of mutuality.

Robert Stalnaker has discussed similar problems [22], 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, $[a][b][a]A$ represents Ann's C-supposition that Bob C-supposes that Ann C-supposes that p. Their mutuality is possible because the participants 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.[6]

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.[7] Here, we find ourselves 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 the same way.

Suppose, for instance, that Ann meets Bob at an AAAI conference, and begins a conversation. She supposes that Bob is a computer scientist. Assume that when Ann learned computer science, she kept track of the things she learned that she could expect any computer scientist to learn. It will be possible for her to do this if, as she learned things, she remembered not only the learned items but the type of circumstance in which they were learned.

In Thomason [24], I propose to model this using a modality $[a, \textsc{cs}]$ to represent the things that Ann expects any computer scientist to believe.[8] Now, Ann expects computer scientists to organize what they learned in much the same way. So she not only expects Bob to have learned, for instance, that finite state automata accept regular languages, but that any computer scientist can be expected to have learned this. That is, a modal model of Ann's mental contents will contain not only $[a, \textsc{cs}]A$, but $[a, \textsc{cs}][b, \textsc{cs}]A$.

Introducing into the makeup of single agents special-purpose modalities whose purpose is mutual modeling provides a more plausible way of producing iterated attitudes. Below, in Section 8, I show how this idea can lead to infinite iteration complexities, and so—at least, under favorable conditions—to full mutuality.

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

[7] See Clark & Schober [6, pp. 257–258].

[8] More generally, if Ann can match Bob to features $M_1 \ldots M_n$, she can use a modality $[M_1 \sqcap \ldots \sqcap M_n]$ to initialize the conversation, where $wR_{M_1 \sqcap \ldots \sqcap M_n}w'$ iff for all i, $1 \leq i \leq n$, wR_iw'.

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 indispensable in accounting in practice for a wide range of examples that involve mutuality.

2.2.3. 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. In a story in which 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 in part to the fact that we tend to imagine normal cases, especially in interpreting narratives.

The plausibility of the examples that are given to support immediacy theories depends to a large extent on the fact that we read them as narratives. In general, we expect authors to state explicitly anything that is abnormal in the circumstances they relate. So, in Schiffer's example, for instance, we assume that the candle is not a novelty item designed to look like a wine bottle. This assumption makes the case for MutAx more plausible.

In real life, we often have to deal with cases that are less 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 initializing assumptions we make that are not based on immediately perceived mutual situations are even more patently defeasible. For instance, Example 2 fails if one participant takes 'here' to refer to a street corner, while the other takes it to refer to a city. Part of being a skilled conversationalist is to make well-coordinated 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.[9]

Axioms like MutAx are inadequate in accounting for such phenomena. To qualify as an axiom, MutAx 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 MutAx, 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

$$\textbf{Qualified MutAx} \qquad [\,a\,][[A \wedge B] \rightarrow [\,\textsc{mut}\,]C],$$

where B 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.[10] Second, we are typically

[9] See Mortensen [19].

[10] This is the qualification problem. See, for instance, McCarthy [16].

willing to use Qualified MutAx 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. The logic represents individual beliefs as interrelated families of modalities; this mechanism provides a workable approach to the problem of initializing mutual attitudes.

3. Subagent simulation as an agent modeling mechanism

From uses of multiple modalities in logical models of multi-agent systems (see Fagin et al. [7]) and contextual reasoning (see Buvač & Mason [3]), we are familiar with 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 AAAI conference example of Section 2.2.2. 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 $[a, \text{CS}]$ 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 & Mason [3]. So the general contours of the theory should be recognizable to those who are familiar with these theories. For instance, there will be "lifting rules" that organize the distribution of information among the subagents of a single agent. Although an agent $\boldsymbol{a}$ can obtain information from another agent $\boldsymbol{b}$ (for instance, by communication), this is not a matter of $\boldsymbol{a}$'s internal epistemic organization, and we certainly do not want to relate indices $\langle a, i\rangle$ and $\langle b, i\rangle$ by monotonic lifting rules. But $\boldsymbol{a}$'s beliefs about $\boldsymbol{b}$'s beliefs do in general depend on beliefs of the subagents of $\boldsymbol{a}$ that imitate subagents of $\boldsymbol{b}$; so we will have lifting rules, rules that relate beliefs of some of $\boldsymbol{a}$'s subagents to $\boldsymbol{a}$'s beliefs about $\boldsymbol{b}$'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. We begin with the intra-agent logic.

4. Modeling the multiplicity of single-agent beliefs[11]

Some subagents can *access* other subagents. This is not a form of communication. It reflects the modular epistemic organization of a single agent. I will not go into details here, but I believe that this 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,

[11] Some of this section corresponds to parts of Thomason [24].

as discussed in Guha [9] 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 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 definition.

Definition 4.3. *Intra-Agent Modal Languages*
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.

$\mathcal{I}$ is the set of subagents of the language, and $\preceq$ determines accessibility for subagents. If $i \preceq j$ then i accesses j.

Definition 4.4. *Intra-Agent Modal Formulas*
The set $\textsc{Formulas}(\mathcal{P}, \mathcal{I})$ is the smallest set that (1) contains $\mathcal{P}$, (2) is closed under boolean connectives, and (3) is closed under i-necessitation, for $i \in \mathcal{I}$. I.e., for all $i \in \mathcal{I}$, if $A \in \textsc{Formulas}(\mathcal{P}, \mathcal{I})$, then $\texttt{[}i\texttt{]}A \in \textsc{Formulas}(\mathcal{P}, \mathcal{I})$.

Definition 4.5. *Intra-Agent Modal Frames.*
An *intra-agent frame* $\mathcal{F} = \langle W, R \rangle$ for an intra-agent modal language $\mathcal{L} = \langle \mathcal{I}, \preceq, \mathcal{P} \rangle$ on a frame W consists of a nonempty set W of possible worlds and a function R providing a relation R_i over W for each $i \in \mathcal{I}$.

Definition 4.6. *Intra-Agent Modal Models.*
A *model interpretation* M on an intra-agent frame $\mathcal{F} = \langle W, R \rangle$ for an intra-agent modal language $\mathcal{L} = \langle \mathcal{I}, \preceq, \mathcal{P} \rangle$ is an assignment of values $[\![P]\!]_M$ to the basic propositions P in $\mathcal{P}$; $[\![P]\!]_M$ is a function from W to $\{T, F\}$. An *intra-agent modal model* $\mathcal{M}$ of an intra-agent modal language $\mathcal{L} = \langle \mathcal{I}, \preceq, \mathcal{P} \rangle$ on a frame $\langle W, R \rangle$ consists of a triple $\langle M, w_0, i \rangle$, where M is a model interpretation on $\mathcal{F}$, $w_0 \in W$, and $i \in \mathcal{I}$. (w_0 is the initial world of the model, i is the designated subagent.)

The satisfaction relation $\mathcal{M} \models_{i,w} A$ is relativized to subagents as well as to worlds; formulas are true or false relative not only to a world, but to a subagent. $\mathcal{M} \models_{i,w} A$ means that $\mathcal{M}$ makes A true in w from the perspective of subagent i. The semantic effects of perspective are very limited; perspective influences only the truth values of modal formulas, and it affects these only in a limited way.

Definition 4.7. *Satisfaction in an Intra-Agent Modal System*
$\mathcal{M} \models_{i,w} P$ iff $[\![P]\!]_M(w) = T$. (Note that i is redundant in this case.) Satisfaction is standard for boolean connectives, and $\mathcal{M} \models_{i,w} \texttt{[}j\texttt{]}A$ iff $i \preceq j$ and for all w such that wR_jw', $\mathcal{M} \models_{j,w'} A$.

Finally, $\mathcal{M} \models A$ iff $\mathcal{M} \models_{i,w_0} A$.

Depending on the application, we may wish to impose certain constraints on the relations R_i. Here, we are interested in the following conditions.

Transitivity. If wR_iw' and $w'R_iw''$ then wR_iw''.

Euclideanness. If wR_iw' and wR_iw'' then $w'R_iw''$.

Seriality. For all w, there is a w' such that wR_iw'.

Subagent Monotonicity. $R_i \subseteq R_j$ if $i \preceq j$.

Subagent Coherence. If wR_iw' and $i \preceq j$ then $w'R_jw'$.

The combination of Transitivity, Euclideanness, and Seriality is commonly used in contemporary logical models of single-agent belief; see Fagin et al. [7]. The remaining constraints are specific to intra-agent epistemic logic.

The resulting logic is a multimodal version of the non-normal logic **E2** that is formulated in Lemmon [13] and proved complete in Kripke [12].

I have a sound and complete axiomatization of the logic; but I will not discuss those details here.

5. Modeling the beliefs of many agents[12]

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. [7].

Intra-agent and multi-agent epistemic logic model fundamentally different aspects of reasoning about attitudes. 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 consulting its opinions.

I will now assume that we have indices for agents as well as for the associated subagents. Thus, we will have formulas like

$$\texttt{[}a, i\texttt{]}[A \rightarrow \texttt{[}b, j\texttt{]}[B \rightarrow \texttt{[}a, i\texttt{]}C]],$$

where a and b are agent indices. Where A, B, and C express p, q, and r, respectively, this formula says that $\boldsymbol{a}$'s i-module believes that if p then $\boldsymbol{b}$'s j-module believes that if q then $\boldsymbol{a}$'s i-module believes that r.

The notation may appear to assume that each subagent knows about how each other agent is decomposed into subagents, but this is required by the fact that this decomposition is built into the language—and we do assume that a uniform language is available for subagents. Depending on the possibilities, any agent's subagents may be well informed or entirely ignorant about the beliefs of the subagents of other agents.

Definition 5.8. *Inter-Agent Modal Languages.*
An inter-agent propositional language $\mathcal{L} = \langle \mathcal{P}, \mathcal{E}, \mathcal{A}, \mathcal{I}, \preceq \rangle$ is determined by a nonempty set $\mathcal{P}$ of predicate letters; by a set $\mathcal{E}$ of individual constants; 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 (representing the subagents of $\boldsymbol{a}$); and by a function $\preceq$ which for each $a \in \mathcal{A}$ provides a reflexive, transitive ordering on $\mathcal{I}_a$. We require that $\mathcal{A} \subseteq \mathcal{E}$—each agent index also

[12] Parts of this section correspond to parts of Thomason [24].

serves as an individual constant.

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.

In circumscriptive theories of belief, we will need abnormality predicates relating agents and propositions; for instance, we may want to characterize a proposition just asserted by $\boldsymbol{a}$ in a conversation as abnormal for $\boldsymbol{a}$ if it is not heard or not understood. So we will want formulas such as $Ab(a, P(b))$. It may be important to record whether agents believe that abnormalities hold, so we will need formulas such as $[\,a, i\,]Ab(c, P(b))$. Although I will not need them in this paper, I will not exclude formulas like $Ab_1(a, Ab_2(b, P))$. These ideas lead to the following formation rules.

A predicate letter P represents a relation over a finite number of arguments, each of which is either an individual or a proposition. The number of arguments may be zero; in that case, P is a propositional constant. I assume that there is at least one non-propositonal predicate letter in $\mathcal{P}$, i.e., at least one predicate letter whose arguments are all individuals. To simplify things, I will also assume that arguments of individual type occur before arguments of propositional type.

Where $\mathcal{L} = \langle \mathcal{P}, \mathcal{E}, \mathcal{A}, \mathcal{I}, \preceq \rangle$, the set $\mathcal{B}_{\mathcal{L}}$ of *basic formulas* of $\mathcal{L}$ consists of all formulas of the form P, where P is a zero-place predicate letter in $\mathcal{P}$, or of the form $P(t_1, \ldots, t_n)$, where P is an n-place nonpropositional predicate letter in $\mathcal{P}$ and $t_1, \ldots, t_n \in \mathcal{E}$.

The set $\textsc{Formulas}(\mathcal{P}, \mathcal{I}, \mathcal{A})$ is the smallest set that:

(1) extends $\mathcal{B}_{\mathcal{L}}$;

(2) is closed under boolean connectives;

(3) is closed under application of basic predicates to complex propositional arguments (i.e., $P(t_1, \ldots, t_n, A_1, \ldots, A_m) \in \textsc{Formulas}(\mathcal{P}, \mathcal{I})$ if $t_1, \ldots, t_n \in \mathcal{E}$, $A_1, \ldots, A_m \in \textsc{Formulas}(\mathcal{P}, \mathcal{I}, \mathcal{A})$, and P takes n individual and m propositional arguments; and

(4) is closed under i, a-necessitation for all $a \in \mathcal{A}$, $i \in \mathcal{I}_a$ (i.e., if $A \in \textsc{Formulas}(\mathcal{P}, \mathcal{I}, \mathcal{A})$, then $[\,a, i\,]A \in \textsc{Formulas}(\mathcal{P}, \mathcal{I})$, for all $a \in \mathcal{A}$ and $i \in \mathcal{I}_a$).

When we speak of a formula $[\,a, i\,]A$, we presuppose that $i \in \mathcal{I}_a$.

Definition 5.9. *Inter-Agent Modal Frames.*
An *inter-agent frame* $\mathcal{F} = \langle W, D \rangle$ for $\mathcal{L} = \langle \mathcal{P}, \mathcal{E}, \mathcal{A}, \mathcal{I}, \preceq \rangle$ consists of nonempty sets W and D. W is the set of possible worlds of the frame and D is the domain of individuals. We will require that $\mathcal{A} \subseteq D$—each agent index is also an individual. This means that we will treat agent indices as self-denoting individual constants.

We use relations on W as usual, to interpret modal operators, but now it makes better sense to include these relations in the models rather than in the frames.

Definition 5.10. *Inter-Agent Modal Models.*
A *model interpretation* M on an inter-agent frame $\mathcal{F} = \langle W, D \rangle$ for a language $\mathcal{L} = \langle \mathcal{P}, \mathcal{E}, \mathcal{A}, \mathcal{I}, \preceq \rangle$ is an assignment of appropriate values, determined by W and D, to constants of $\mathcal{L} = \langle \mathcal{P}, \mathcal{E}, \mathcal{A}, \mathcal{I}, \preceq \rangle$, where here we include modalities among the constants. In particular,

1. $[\![t]\!]_M \in D$ for individual constants t; and if $t \in \mathcal{A}$, $[\![t]\!]_M = t$.
2. If P is a predicate letter taking n individual arguments and m propositional arguments, then $[\![P]\!]_M$ is assigned a function from W to subsets of $D^n \times \mathcal{P}(W)^m$, where $\mathcal{P}(W)$ is the power set of W. (Thus, for each world this function delivers an appropriate relation over individuals and propositions.)
3. $[\![\,\texttt{[}\,a, i\,\texttt{]}\,]\!]_M$ is a transitive, Euclidean, serial relation on W.

A *model* on an intra-agent frame $\mathcal{F} = W$ for a language $\mathcal{L} = \langle \mathcal{P}, \mathcal{E}, \mathcal{A}, \mathcal{I}, \preceq \rangle$ is a tuple $\mathcal{M} = \langle M, w_0, a, i \rangle$, where M is a model interpretation, $w_0 \in W$, $a \in \mathcal{A}$, and $i \in \mathcal{I}_a$. (For many purposes, we can neglect the role that a and i play in satisfaction—often, then, we can simply think of a model as a pair $\langle M, w_0 \rangle$.) Where δ is a constant of $\mathcal{L}$, we let $[\![\delta]\!]_{\mathcal{M}} = [\![\delta]\!]_M$.

Definition 5.11. *Satisfaction.*
The satisfaction relation $\mathcal{M} \models_{a,i,w} A$ is relativized to subagents and to worlds, as before. (But the two indices a, i reflect the fact that a subagent is now a pair consisting of an agent and an index.)

For formulas A of the form $P(t_1, \ldots, t_n)$, $\mathcal{M} \models_{a,i,w} A$ iff $\langle [\![t_1]\!]_{\mathcal{M}}, \ldots, [\![t_n]\!]_{\mathcal{M}} \rangle \in [\![P]\!]_{\mathcal{M}}(w)$.

For any formula A, $[\![A]\!]_{\mathcal{M},a,i} = \{w \in W : \ \mathcal{M} \models_{a,i,w} A\}$.

For formulas A of the form $P(t_1, \ldots, t_n, A_1, \ldots, A_m)$, $\mathcal{M} \models_{a,i,w} A$ iff $\langle [\![t_1]\!]_{\mathcal{M}}, \ldots, [\![t_n]\!]_{\mathcal{M}}, [\![A_1]\!]_{\mathcal{M},a,i}, \ldots, [\![A_m]\!]_{\mathcal{M},a,i} \rangle \in [\![P]\!]_{\mathcal{M}}(w)$.

Satisfaction conditions are standard for boolean connectives, and finally

$\mathcal{M} \models_{a,i,w} \texttt{[}\,b, j\,\texttt{]}\,A$ iff either (1) $a = b$, $i \preceq_a j$, and $\mathcal{M} \models_{b,j,w'} A$ for all w' such that $wR_{a,j}w'$, or (2) $a \neq b$ and $\mathcal{M} \models_{a,i,w'} A$ for all w' such that $wR_{b,j}w'$.

Where $\mathcal{M} = \langle M, w_0, a, i \rangle$ is a model, $\mathcal{M} \models A$ iff $\mathcal{M} \models_{a,i,w_0} A$. Where T is a set of formulas, $\mathcal{M}$ *simultaneously satisfies* T, or *is a model of* T, iff $\mathcal{M} \models A$ for all $A \in T$. And T *implies* A iff $\mathcal{M} \models A$ for all models $\mathcal{M}$ that simultaneously satisfy T.

It is reasonable to require that for all $a \in \mathcal{A}$ there is a unique $i_a \in \mathcal{I}_a$ that is $\preceq_a$ minimal in $\mathcal{I}_a$: for all $j \in \mathcal{I}_a$, $i \preceq_a j$; i_a represents the compiled beliefs of agent $\boldsymbol{a}$.

As before, we are primarily interested in models that are transitive, Euclidean, and serial. Although I am also interested in nonmonotonic relaxations of subagent monotonicity, so far I have only looked at the case in which models are subagent monotonic and coherent.

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 $\mathcal{A} = \{a\}$.

For the remainder of this paper, I will only consider *agent-homogeneous* languages, where $\mathcal{I}_a = \mathcal{I}_b$ for all a and b. In this case, for each $i \in \mathcal{I}$ we can add a mutual belief operator $\texttt{[}\,\textsc{mut}, i\,\texttt{]}$ to the logic. We can also simplify things by thinking of a language $\mathcal{L}$ as a tuple $\langle \mathcal{P}, \mathcal{E}, \mathcal{A}, \mathcal{I}, \preceq \rangle$, where $\mathcal{I}$ is a fixed set of indices. From now on, I will work with this simpler account of languages.

Definition 5.12. *Inter-Agent Modal Systems with Mutual Belief.*
An inter-agent propositional language $\mathcal{L} = \langle \mathcal{P}, \mathcal{E}, \mathcal{A}, \mathcal{I}, \preceq, \textsc{mut}\rangle$ with mutual belief is an agent-homogeneous inter-agent propositional language with a modal operator $\texttt{[}\,\textsc{mut}, i\,\texttt{]}$ for each $i \in \mathcal{I}$. The satisfaction condition for $\texttt{[}\,\textsc{mut}, j\,\texttt{]}$ is as follows:

> $\mathcal{M} \models_{a,i,w} \texttt{[}\,\textsc{mut}, j\,\texttt{]}A$ iff $i \preceq j$ and $\mathcal{M} \models_{a,i,w'} A$ for all w' such that $wR^*_j w'$, where R^*_j is the transitive closure of the set of relations $\{R_{b,j} : b \in \mathcal{A}\}$.

The resulting logic contains standard multi-agent modal logics for reasoning about mutual belief, such as the system $\mathbf{KD45}^C_n$ of Fagin et al. [7].

Definition 5.13. $\texttt{[}\,\alpha, i\,\texttt{]}$
Where α is the string $\alpha_1 \ldots \alpha_n$ over $\mathcal{A}$, we let $\texttt{[}\,\alpha, i\,\texttt{]} = \texttt{[}\,\alpha_1, i\,\texttt{]} \ldots \texttt{[}\,\alpha_n, i\,\texttt{]}$.

Lemma 5.1. $\{\texttt{[}\,\alpha, i\,\texttt{]}A : \alpha$ a nonempty string over $\mathcal{A}\}$ implies $\texttt{[}\,\textsc{mut}, i\,\texttt{]}A$.
Proof. Suppose that $\mathcal{M} = \langle M, w_0, a, j\rangle$ is a model for $\mathcal{L} = \langle \mathcal{P}, \mathcal{E}, \mathcal{A}, \mathcal{I}, \preceq, \textsc{mut}\rangle$ that simultaneously satisfies $\{\texttt{[}\,\alpha, i\,\texttt{]}A : \alpha$ a string over $\mathcal{A}\}$, and let $w_0 R^*_i w$. Then for some $\alpha = \alpha_1 \ldots \alpha_n$ and $w_1, \ldots, w_n$, we have $w_0 R_{\alpha_1,i}\, w_1, \ldots, w_{n-1} R_{\alpha_n,i}\, w_n$, where $w_n = w$. By hypothesis, $\mathcal{M} \models_{a,j,w_0} \texttt{[}\,\alpha, i\,\texttt{]}A$. So $\mathcal{M} \models_{a,j,w} A$. Now, we also have $j \preceq i$, since $\mathcal{M} \models_{a,j,w_0} \texttt{[}\,a, i\,\texttt{]}A$. Therefore, $\mathcal{M} \models_{a,j,w} \texttt{[}\,\textsc{mut}, i\,\texttt{]}A$.

Lemma 5.2. $\{\texttt{[}\,\alpha, j\,\texttt{]}\texttt{[}\,\beta, i\,\texttt{]}A : \beta$ a nonempty string over $\mathcal{A}\}$ implies $\texttt{[}\,\alpha, j\,\texttt{]}\texttt{[}\,\textsc{mut}, i\,\texttt{]}A$.
Proof. This follows directly from Lemma 1, together with principles of modal logic.

Lemma 5.3. $\{\texttt{[}\,a\,\texttt{]}\texttt{[}\,\textsc{mut}\,\texttt{]}A : a \in \mathcal{A}\}$ implies $\texttt{[}\,\textsc{mut}\,\texttt{]}A$.
Proof. Clearly, $\{\texttt{[}\,a\,\texttt{]}\texttt{[}\,\textsc{mut}\,\texttt{]}A : a \in \mathcal{A}\}$ implies $\texttt{[}\,\alpha\,\texttt{]}A$ for all nonnull strings α over $\mathcal{A}$, since any such string has the form $b\alpha$, where $b \in \mathcal{A}$. By Lemma 1, then, $\{\texttt{[}\,a\,\texttt{]}\texttt{[}\,\textsc{mut}\,\texttt{]}A : a \in \mathcal{A}\}$ implies $\texttt{[}\,\textsc{mut}\,\texttt{]}A$.

Together, these lemmas provide us with a methods for establishing mutuality and attitudes about mutuality. These methods will be used below in Section 8 to show how beliefs about mutuality (and, in favorable circumstances, mutuality itself) can be secured by default.

6. Formulating propositional generalizations

In formalizing the problem of achieving mutual attitudes, we will need to express generalities about agents and propositions. A simple generalization of the sort I have in mind would be an axiom saying of a particular "bellwether" agent **a** that if this agent's public module believes anything, then every agent's public module believes this thing. It is most natural to use propositional quantifiers to state such generalizations:

(6.1) $\forall P[\texttt{[}\,a, \textsc{pub}\,\texttt{]}P \rightarrow \forall x[\texttt{Agent}(x) \rightarrow \texttt{[}\,x, \textsc{pub}\,\texttt{]}P]]$

This axiom uses a quantifier $\forall x$ over individuals and a quantifier $\forall P$ over propositions.

Unfortunately, propositional quantifiers complicate the logical setting. In cases where there is only one modal operator, the complexity of logics with propositional quantification differs wildly, depending on the type of the modality. The logic could be decidable, or it could be equivalent to second-order modal logic. (See Fine [8] and Kremer [11].) As far as I know, the case with multiple modalities has not been investigated to any extent, but I have an unpublished proof that, even with the modalities that are decidable in the monomodal case, multimodal logics with propositional quantifiers are equivalent to full second-order logic. I believe that this complexity is an inevitable consequence of a general approach to the phenomenon of knowledge transfer.[13] Hopefully, however, tractable special cases will emerge.

Worse, propositional quantifiers create special difficulties in combining epistemic logic with circumscription, by making the technique of model separation introduced in Section 7 unusable. I have a way to overcome this problem using nonstandard models, but I can spare you the details of this by taking advantage of the fact that the only constraints needed to secure mutuality have the form

(6.2) $\forall P_1 \ldots \forall P_n A$,

where A has no propositional quantifiers. This means that axiom schemata can be used in place of axioms that involve propositional quantification. The use of axiom schemata means that the theories we will consider will not be finitely axiomatizable, but that is a pretty small price to pay in this context.

By confining ourselves to cases where there are finitely many agents, we can (temporarily, at least) dispense with quantification over individuals. Our investigation of the logic of mutuality, then, will take place in the context of inter-agent modal propositional logics with mutual belief.

7. Circumscriptive Reasoning about Beliefs

Several frameworks have been proposed for formalizing nonmonotonic reasoning about beliefs: autoepistemic logic, Morgenstern [18]; circumscription in a higher-order modal logic, Thomason [23]; default logic (or something like it), Parikh [20]; only-knowing, Halpern & Lakemeyer [10]; and preferential models, Wainer [25] Monteiro & Wainer [17].

I will adopt a circumscriptive approach, partly because it is useful and straightforward as a development tool, and partly because circumscription appears to provide more powerful and flexible mechanisms for describing complex configurations of abnormalities. Fortunately, the interrelations between the various approaches to nonmonotonic logic are by now pretty thoroughly worked out, and in the simple cases at least it is possible to go fairly easily from one to the other. Hopefully this will carry over to epistemic applications.

The early circumscriptive accounts of nonmonotonic reasoning appeal to a completion operation on finitely axiomatized theories, which takes the original theory into one in which certain terms are minimized. However, later formulations dispense with the need for a finitely axiomatized theory by defining circumscriptive consequence in model theoretic terms. Here, I will use the latter approach.

Circumscriptive modal logics are a neglected topic. In thinking through this application of circumscription, I encountered technical difficulties, due to the interaction of the possible

[13] Similar problems arise, for instance, in the general logic of contextual reasoning.

worlds interpretation of propositional attitudes with circumscription. The purpose of the next two paragraphs is to explain these difficulties and to motivate their formal solution.

In circumscribing a theory, we minimize the extensions of some constants in a space of models in which the extensions of other constants are allowed to vary. In standard circumscriptive theories, the constants are all either first-order predicates or functors; but there are no formal difficulties in applying circumscriptive techniques to constants of other logical types, and in particular to modal operators, which for our purposes can be thought of as predicates of propositions. In fact, since we are interested in minimizing differences between the beliefs of agents, the only constants that are minimized or varied in simple cases will be abnormality predicates (which mark discorrelations among certain beliefs) and modalities. (In the presence of a theory of abnormalities, it may be necessary to consider more complex patterns of variation.) In possible worlds semantics, beliefs are not represented directly as predicates of propositions but are characterized indirectly in terms of relations over possible worlds. This lack of uniformity creates some difficulties in motivating an appropriate account of circumscription.

Suppose that we are trying to minimize certain abnormalities for an agent $\boldsymbol{a}$, in order to make $\boldsymbol{a}$'s picture of $\boldsymbol{b}$'s beliefs correspond, as far as possible, to $\boldsymbol{a}$'s own beliefs. Intuitively, the things we want to vary here are $\boldsymbol{a}$'s beliefs about $\boldsymbol{b}$'s beliefs. But in all our models, the only parameter we can vary is the relation $R_{a,i}$ that serves to interpret the modal operator $\texttt{[}\,a, i\,\texttt{]}$ in which we are interested. Suppose we are working with a base world w_0. The closest we can come to varying $\boldsymbol{a}$'s beliefs about $\boldsymbol{b}$'s beliefs would be to vary how $\boldsymbol{b}$'s beliefs work in worlds epistemically accessible for $\boldsymbol{a}$ from w_0. That is, we vary the worlds b, i-related to w, for worlds w such that $w_0 R_{a,i} w$. The problem with this, however, is that there is no guarantee that such changes will be confined to $\boldsymbol{a}$'s beliefs about $\boldsymbol{b}$. We can easily have both $w_0 R_{a,i} w$ and $w_0 R_{b,i} w$. In this case, if we change $W_{b,i,w}$ we are also changing $\boldsymbol{b}$'s beliefs. But intuitively, when we vary $\boldsymbol{a}$'s beliefs about $\boldsymbol{b}$, we want to hold $\boldsymbol{b}$'s actual beliefs constant. So we need to exclude cases in which for some w, $w_0 R_{a,i} w$ and $w_0 R_{b,i} w$.

More generally, we need to look at chains of related worlds originating at w_0. That is, we need to ensure that $W^*_{w_0,a,i}$ and $W^*_{w_0,b,j}$ are disjoint, where $W^*_{w_0,a,i}$ and $W^*_{w_0,b,j}$ are defined below in Definition 7.

Definition 7.14. $W_{w,\alpha,i}$ *(where* $\alpha \in \{a, b\}$*).* $W_{w,\alpha,i} = \{w' : wR_{\alpha,i}w'\}$.

Definition 7.15. $W^*_{w,\alpha,i}$
$W^*_{w,\alpha,i}$ is the smallest subset of W such that (1) $W_{\alpha,i} \subseteq W^*_{w,\alpha,i}$ and (2) if $w_1 \in W^*_{w,\alpha,i}$ and $w_1 R_{\beta,j} w_2$ for any agent-subagent pair $\langle \beta, j\rangle$, where $\beta \in \mathcal{A}, \beta \neq \alpha$, then $w_2 \in W^*_{w,\alpha,i}$.

We arrive at the following definition of a *separated* model by applying the disjointness requirement to all differing agent-subagent pairs; we also require that $w_0 \notin W^*_{w_0,a,i}$ for all agent-subagent pairs $\langle a, i\rangle$.

Definition 7.16. *Separated model.*
Let $\mathcal{M} = \langle M, w_0, i, a\rangle$ be a model on a frame $\mathcal{F} = \langle W, S\rangle$ for a language $\mathcal{L} = \langle \mathcal{P}, \mathcal{E}, \mathcal{A}, \mathcal{I}, \preceq\rangle$. $\mathcal{M}$ is *separated* iff (1) for all agent-subagent pairs $\langle b, j\rangle, \langle b', j'\rangle \in \mathcal{A} \times \mathcal{I}$ such that $\langle b, j\rangle \neq \langle b', j'\rangle$, $W^*_{w_0,b,j} \cap W^*_{w_0,b',j'} = \emptyset$ and (2) $w_0 \notin W^*_{w_0,b,j}$ for all agent-subagent pairs $\langle b, j\rangle$.

It is easy to show that restricting our attention to separated models will lose no generality; for every model, there is a separated model that satisfies exactly the same formulas. The construction that proves this fact involves "cloning" worlds, replacing each $w \in W^*_{w_0,b,j}$ with a copy $\langle w, b, j\rangle$ indexed to $\langle b, j\rangle$.

Theorem 7.1. For every model $\mathcal{M} = \langle M, w_0, a, i\rangle$ of an inter-agent propositional language $\mathcal{L}$ with mutual belief on a frame $\mathcal{F} = \langle W, D\rangle$, there is a separated model $\mathcal{M}' = \langle M', w_0, a, i\rangle$ of $\mathcal{L}$ on $\mathcal{F}' = \langle W', D\rangle$ such that for all formulas A of $\mathcal{L}$, $\mathcal{M} \models A$ iff $\mathcal{M}' \models A$.

As I noted above in Section 6, this result fails, for standard models at least, if propositional quantifiers are added.

We are now in a position to define the apparatus that, using circumscription, will provide a nonmonotonic consequence relation. Suppose that we have a group $\mathcal{G}$ of agents, and we are interested in the beliefs of the members of some subgroup $\mathcal{G}'$ of $\mathcal{G}$ about the beliefs of the members of $\mathcal{G}$. To locate the appropriate beliefs, we fix some subagent index i; in general, this will be an index that is taken to be public for $\mathcal{G}$ by the members of $\mathcal{G}$. Finally, suppose that we are working with a set $\boldsymbol{Ab}$ of abnormality predicates which we wish to minimize, with respect to the beliefs of the i-subagents of members of $\mathcal{G}'$ about the members of $\mathcal{G}$.

We begin with a separated model $\mathcal{M}$. First, we identify a class of models (not necessarily separated) obtained from $\mathcal{M}$ by allowing abnormalities and the beliefs of the i-subagents of $\mathcal{G}'$ about the beliefs of other members of $\mathcal{G}$ to vary freely. Within this class, we prefer models with fewer abnormalities. We then define the formulas circumscriptively implied by a theory T by restricting our attention to minimal models of T, relative to this preferential ordering. All this is made precise in the following definitions.

Definition 7.17. $\mathcal{M}_1 \cong_{\boldsymbol{Ab},\mathcal{G},\mathcal{G}',j} \mathcal{M}_2$.
Let $\mathcal{G}' \subseteq \mathcal{G} \subseteq \mathcal{A}$. Let $\mathcal{M}_1 = \langle M_1, w_0, a, i\rangle$ and $\mathcal{M}_2 = \langle M_2, w_0, a, i\rangle$ be separated models on the same frame such that M_1 and M_2 are alike except on a set $\boldsymbol{Ab}$ of predicate constants and on the set of modalities $\{[\,b, j\,] : b \in \mathcal{G}'\}$. Let $R^1_{b,k} = [\![[\,b, k\,]]\!]_{\mathcal{M}_1}$, $R^2_{b,k} = [\![[\,b, k\,]]\!]_{\mathcal{M}_2}$. Then $\mathcal{M}_1 \cong_{\boldsymbol{Ab},\mathcal{G},\mathcal{G}',j} \mathcal{M}_2$ iff the following holds: $wR^1_{b,k}w'$ iff $wR^2_{b,k}w'$ unless for some $a \in \mathcal{G}'$, $w_0R^1_{a,j}w$, $j = k$, $b \neq a$, and $b \in \mathcal{G}$.

Definition 7.18. $\mathcal{M}_1 \leq_{\boldsymbol{Ab}} \mathcal{M}_2$.
Let $\mathcal{M}_1 = \langle M_1, w_0, a, i\rangle$ and $\mathcal{M}_2 = \langle M_2, w_0, a, i\rangle$ be models on the same frame. Let $\boldsymbol{Ab}$ be a set of predicate constants. Then $\mathcal{M}_1 \leq_{\boldsymbol{Ab}} \mathcal{M}_2$ iff for all $P \in \boldsymbol{Ab}$, $[\![P]\!]_{\mathcal{M}_1}(w_0) \subseteq [\![P]\!]_{\mathcal{M}_2}(w_0)$.

Definition 7.19. $\mathbf{Ab}, \mathcal{G}, \mathcal{G}', i$*-minimality for* T.
Let T be a set of formulas of $\mathcal{L} = \langle \mathcal{B}, \mathcal{I}, \mathcal{A}, \preceq, \textsc{mut}\rangle$, and $\mathcal{M} = \langle M, w_0, a, j\rangle$ be a model of T. $\mathcal{M}$ is $\mathbf{Ab}, \mathcal{G}, \mathcal{G}', i$-minimal for T iff (1) for some separated model $\mathcal{M}_1 = \langle M_1, w_0, a, j\rangle$, $\mathcal{M} \cong_{\boldsymbol{Ab},\mathcal{G},\mathcal{G}',i} \mathcal{M}_1$ and (2) for all models $\mathcal{M}' = \langle M', w_0, a, j\rangle$ of T such that $\mathcal{M}' \cong_{\boldsymbol{Ab},\mathcal{G},\mathcal{G}',i} \mathcal{M}_1$, $\mathcal{M} \leq_{\boldsymbol{Ab}} \mathcal{M}'$.

Definition 7.20. $\|\!\!\sim_{\boldsymbol{Ab},\mathcal{G},\mathcal{G}',i}$.
Let $\mathcal{G}' \subseteq \mathcal{G} \subseteq \mathcal{A}$. Let T be a set of formulas of $\mathcal{L} = \langle \mathcal{B}, \mathcal{I}, \mathcal{A}, \preceq, \textsc{mut}\rangle$. Then $T \|\!\!\sim_{\boldsymbol{Ab},\mathcal{G},\mathcal{G}',\mathrm{i}} A$ iff $\mathcal{M} \models A$ for all models $\mathcal{M} =$ that are $\mathbf{Ab}, \mathcal{G}, \mathcal{G}', i$-minimal for T.

8. Achieving mutuality through nonmonotonic reasoning

8.1. Simplifications

We begin by simplifying things. First, let's assume that there are only two agents. Second, the language is not only homogeneous, but in fact each agent has only one minimal subagent i_0 (which is also public). So there are only two agent modalities: $[\,a, i_0\,] = [\,a, \textsc{pub}\,]$ and $[\,b, i_0\,] = [\,b, \textsc{pub}\,]$. And we confine ourselves to frames with $D = \mathcal{A}$, i.e. the only individuals are agents.

We will only be interested in two abnormality predicates: the predicates Ab^a_1 and Ab^b_1 given in Section 8.2, below. And we will only want to circumscribe in three ways: (1) minimizing Ab^a_1 while allowing $\boldsymbol{a}$'s beliefs about $\boldsymbol{b}$ to vary, (2) minimizing Ab^b_1 while allowing $\boldsymbol{b}$'s beliefs about $\boldsymbol{a}$ to vary, and (3) minimizing Ab^a_1 and Ab^b_1 while allowing $\boldsymbol{a}$'s beliefs about $\boldsymbol{b}$ and $\boldsymbol{b}$'s beliefs about $\boldsymbol{a}$ to vary. In case (1), we are circumscribing only with respect to $\boldsymbol{a}$'s beliefs about $\boldsymbol{b}$; in case (2), we are circumscribing only with respect to $\boldsymbol{b}$'s beliefs about $\boldsymbol{a}$; in case (3), we are circumscribing with respect to $\boldsymbol{a}$'s and $\boldsymbol{b}$'s beliefs about each other.

This enables us to simplify our notation for circumscription. The following definitions provide for the three types of circumscription in which we are interested.

Definition 8.21. *$\mathcal{M}_1 \cong_a \mathcal{M}_2$, $\mathcal{M}_1 \leq_a \mathcal{M}_2$, $\boldsymbol{a}$-minimality for T, $\|\!\!\sim_{\mathrm{a}}$.*
Let
$\boldsymbol{Ab} = \{Ab^a_1, Ab^a_2\}$, $\mathcal{G}' = \{a\}$, $\mathcal{G} = \{a, b\}$, and $i = i_0$. Then:

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

Definition 8.22. *$\mathcal{M}_1 \cong_b \mathcal{M}_2$, $\mathcal{M}_1 \leq_b \mathcal{M}_2$, $\leq_b$-minimality for T , $\|\!\!\sim_{\mathrm{b}}$.*
Let
$\boldsymbol{Ab} = \{Ab^b_1, Ab^b_2\}$, $\mathcal{G}' = \{b\}$, $\mathcal{G} = \{a, b\}$, and $i = i_0$. Then:

(1) $\mathcal{M}_1 \cong_b \mathcal{M}_2$ iff $\mathcal{M}_1 \cong_{\boldsymbol{Ab},\mathcal{G},\mathcal{G}',i} \mathcal{M}_2$;
(2) $\mathcal{M}_1 \leq_b \mathcal{M}_2$ iff $\mathcal{M}_1 \leq_{\boldsymbol{Ab}} \mathcal{M}_2$;
(3) $\mathcal{M}$ is $\boldsymbol{b}$-minimal for T iff $\mathcal{M}$ is $\mathbf{Ab}, \mathcal{G}, \mathcal{G}', i$-minimal for T;
(4) $T \|\!\!\sim_{\mathrm{b}} A$ iff $T \|\!\!\sim_{\boldsymbol{Ab},\mathcal{G},\mathcal{G}',\mathrm{i}} A$.

Definition 8.23. *$\mathcal{M}_1 \cong_{a,b} \mathcal{M}_2$, $\mathcal{M}_1 \leq_{a,b} \mathcal{M}_2$, $\leq_{a,b}$-minimality for T , $\|\!\!\sim_{\mathrm{a,b}}$.*

Let $\boldsymbol{Ab} = \{Ab^a_1, Ab^a_2, Ab^b_1, Ab^b_2\}$, $\mathcal{G}' = \mathcal{G} = \{a, b\}$, and $i = i_0$. Then:

(1) $\mathcal{M}_1 \cong_{a,b} \mathcal{M}_2$ iff $\mathcal{M}_1 \cong_{\boldsymbol{Ab},\mathcal{G},\mathcal{G}',i} \mathcal{M}_2$;
(2) $\mathcal{M}_1 \leq_{a,b} \mathcal{M}_2$ iff $\mathcal{M}_1 \leq_{\boldsymbol{Ab}} \mathcal{M}_2$;
(3) $\mathcal{M}$ is a, b-minimal for T iff $\mathcal{M}$ is $\mathbf{Ab}, \mathcal{G}, \mathcal{G}', i$-minimal for T;
(4) $T \mathrel{\|\!\!\sim}_{\mathrm{a,b}} A$ iff $T \mathrel{\|\!\!\sim}_{\boldsymbol{Ab},\mathcal{G},\mathcal{G}',\mathrm{i}} A$.

We can also simplify models. In this case, where there is only one subagent per agent, satisfaction in a model does not depend on a choice of any particular subagent. So instead of treating models as structures $\langle M, w_0, a, i\rangle$, we can simplify to $\langle M, w_0\rangle$.

Together, these simplifications make it much easier to present the basic results; and I do not believe that they lose any significant generality.

8.2. 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.

$$(\mathrm{Tr}^a_1)\quad [[a]A \wedge \neg Ab^a_1(A)] \rightarrow [a][b]A$$
$$(\mathrm{Tr}^a_2)\quad [[a][b]A \wedge \neg Ab^a_2(A)] \rightarrow [a]A$$
$$(\mathrm{Tr}^b_1)\quad [[b]A \wedge \neg Ab^b_1(A)] \rightarrow [b][a]A$$
$$(\mathrm{Tr}^b_2)\quad [[b][a]A \wedge \neg Ab^b_2(A)] \rightarrow [b]A$$

According to Axiom Schema (Tr^a_1), $\boldsymbol{a}$ normally believes that $\boldsymbol{b}$ believes whatever $\boldsymbol{a}$ believes; Axiom Schema (Tr^b_1) says the corresponding thing about $\boldsymbol{b}$. The converse axiom schemata (Tr^a_2) and (Tr^b_2) are not needed to show that mutuality (or belief in mutuality) can be obtained in simple cases, but can be useful in formalizing more complex cases. (For an example, see the discussion of Case 3 in Section 9, below.)

The interpretation of the transfer schemata is potentially confusing, and needs to be clarified. I do not require that these schemata are believed; it is compatible with the schemata, for instance, that for some formulas A, $\langle a\rangle[[a]A \wedge \neg Ab^a_1(A) \wedge \neg[a][b]A]$ is true. The transfer schemata require that an abnormality condition must fail for their conclusions to follow; they do not require that the condition should be *believed* to fail. All this makes sense if we think of the transfer schemata not as internal principles that the agents themselves reason with, but as external constraints imposed by an agent designer on the epistemic makeup of agents. The case in which agents are themselves able to reason about abnormalities and transfer constraints is certainly worth considering, but it is more complicated, and the transfer schemata I consider here do not provide a logical model of this reflective case.

I now show that if the extension of $Ab^a_1(A)$ is empty for all A, Axiom Schema (Tr^a_1) implies that whatever $\boldsymbol{a}$ believes, $\boldsymbol{a}$ also believes to be mutual. (And similarly, of course, for Axiom Schema (Tr^b_1) and $\boldsymbol{b}$.) I will use 'Tr^a_1', etc. to denote not only the schemata given above, but the corresponding sets of formulas.

8.3. A result about mutual belief

Let $\mathrm{Normal}_{a,1} = \{\neg Ab_1^a(A) :\ A \text{ a formula}\}$. Consider the following rule concerning mutual belief.

NMB From $\mathrm{Tr}_1^a \cup \mathrm{Normal}_{a,1}$ infer $[\,a\,]A \rightarrow [\,a\,][\,\textsc{mut}\,]A$

Lemma 8.4. The rule (**NMB**) is valid.
Proof. Suppose that $\mathcal{M}$ satisfies Tr_1^a, $\mathrm{Normal}_{a,1}$, and $[\,a\,]A$.

We show by induction on length of α that for all strings α over $\{a, b\}$, $\mathcal{M} \models [\,a\alpha\,]A$. By hypothesis, $\mathcal{M} \models [\,a\alpha\,]A$ when $\alpha = \phi$. Suppose that $\mathcal{M} \models [\,a\alpha\,]A$. Note that $\mathrm{Tr}_1^a \cup \mathrm{Normal}_{a,1}$ implies $[\,a\,]B \rightarrow [\,ab\,]B$, for all formulas B. So, in particular, $\mathrm{Tr}_1^a \cup \mathrm{Normal}_{a,1}$ implies $[\,a\alpha\,]A \rightarrow [\,ab\alpha\,]A$. So $\mathcal{M} \models [\,ab\alpha\,]A$. Also, by the modal logic of $[\,a\,]$, $\mathcal{M} \models [\,aa\alpha\,]A$. This completes the induction. It now follows from Lemma 2 that $\mathcal{M} \models [\,a\,][\,\textsc{mut}\,]p$.

We now show how Lemma 4 allows mutuality to be secured by default.

8.4. Mutuality by default

The theorems in this section show that the transfer schemata ensure that, under fairly 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 $\boldsymbol{a}$ and $\boldsymbol{b}$, 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 $\boldsymbol{a}$ is a morning person, and believes that the lecture will be given at 9am, while $\boldsymbol{b}$ 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 each other 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 will not attempt to formalize these here.

The following result establishes moderately general conditions under which agents satisfying the local epistemic transfer axiom schemata will believe (publicly) that their public beliefs are mutual.[14]

[14] The conditions are actually not general enough to cover many realistic cases. In particular, the restriction on T_2 to formulas containing no occurrences of Ab_1^a or of Ab_2^a makes the result inapplicable to theories that contain plausible axioms about abnormalities. But the technique used to prove the result seems to generalize, and I believe it can be used to cover many realistic cases ininvolving an abnormality theory. I do not yet have a general result covering a suitably large class of these cases.

Lemma 8.5. Let $T = T_1 \cup T_2 \cup T_3 \cup T_4 \cup T_5 \cup T_6 \cup T_7$, where:

1. T_1 consists of all instances of the epistemic transfer axiom schemata for agent $\boldsymbol{a}$: $T_1 = \mathrm{Tr}^a_1 \cup \mathrm{Tr}^a_2$.
2. T_2 consists of boolean formulas (i.e., of formulas that contain no modal operators) that contain no occurrences of Ab^a_1 or of Ab^a_2.
3. T_3 consists of formulas having the form $\mathtt{[}\,a\,\mathtt{]}A$, where A is boolean.
4. T_4 consists of formulas having the form $\mathtt{<}\,a\,\mathtt{>}A$, where A is boolean.
5. T_5 consists of formulas having the form $\mathtt{[}\,b\,\mathtt{]}A$, where A is boolean.
6. T_6 consists of formulas having the form $\mathtt{<}\,b\,\mathtt{>}A$, where A is boolean.
7. T_7 consists of formulas having the form $\mathtt{[}\,a\,\mathtt{]}\textsc{mut}A$, where A is boolean.

Then $T \mathrel{\|\!\!\!\sim_{\mathrm{a}}} \neg Ab^a_1(A)$ for all formulas A.

Proof. We show that for all separated models $\mathcal{M} = \langle M, w_0\rangle$ of T on frame $\mathcal{F} = \langle W, D\rangle$ there is a model $\mathcal{M}' = \langle M', w_0\rangle$ on $\mathcal{F} = \langle W, D\rangle$ such that $\mathcal{M}$ simultaneously satisfies T, $\mathcal{M}' \cong_a \mathcal{M}$ and $[\![Ab^a_1]\!]_{\mathcal{M}'}(w_0) = \emptyset$.

Let $R_a = [\![\mathtt{[}\,a, \textsc{pub}\,\mathtt{]}]\!]_{\mathcal{M}}$ and $R_b = [\![\mathtt{[}\,b, \textsc{pub}\,\mathtt{]}]\!]_{\mathcal{M}}$. Similarly, $R'_a = [\![\mathtt{[}\,a, \textsc{pub}\,\mathtt{]}]\!]_{\mathcal{M}'}$ and $R'_b = [\![\mathtt{[}\,b, \textsc{pub}\,\mathtt{]}]\!]_{\mathcal{M}'}$. And let $W_{w,a} = W_{a,i_a} = \{w' : \ wR_{a,i_a}w'\}$, $W_{w,b} = W_{b,i_b} = \{w' : \ wR_{b,i_b}w'\}$. Define $\mathcal{M}'$ by (i) letting $w_1 R_b w_2$ iff $w_1 R_a w_2$ for $w_1 \in W^a_{w_0}$, (ii) letting $[\![Ab^a_1]\!]_{\mathcal{M}'}(w_0) = [\![Ab^a_2]\!]_{\mathcal{M}'}(w_0) = \emptyset$, and letting $\mathcal{M}'$ agree with $\mathcal{M}$ elsewhere. (What we are doing here is replacing $\boldsymbol{a}$'s beliefs about $\boldsymbol{b}$ with $\boldsymbol{a}$'s beliefs about $\boldsymbol{a}$'s own beliefs.)

Remark (i). For all boolean formulas B not containing occurrences of Ab^a_1 or Ab^a_2, $\mathcal{M} \models_{w_0} B$ if and only if $\mathcal{M}' \models_{w_0} B$.

Remark (ii). For all boolean formulas B and for all $w \in W$, if $w \neq w_0$ then $\mathcal{M} \models_w B$ if and only if $\mathcal{M}' \models_w B$.

It is straightforward to prove these by induction on the complexity of B. The proofs make essential use of the assumption that $\mathcal{M}$ is separated.

We now establish the following claims about $\mathcal{M}'$:

(1.1) $\mathcal{M}'$ simultaneously satisfies Tr^a_1.

(1.2) $\mathcal{M}'$ simultaneously satisfies Tr^a_2.

(2) $\mathcal{M}'$ simultaneously satisfies T_2.

(3) $\mathcal{M}'$ simultaneously satisfies T_3.

(4) $\mathcal{M}'$ simultaneously satisfies T_4.

(5) $\mathcal{M}'$ simultaneously satisfies T_5.

(6) $\mathcal{M}'$ simultaneously satisfies T_6.

(7) $\mathcal{M}'$ simultaneously satisfies T_7.

Proof of (1.1). Let $A \in \mathrm{Tr}^a_1$. Then for some B, A is $[\mathtt{[}\,a\,\mathtt{]}B \wedge \neg Ab^a_1(B)] \rightarrow \mathtt{[}\,a\,\mathtt{]}\mathtt{[}\,b\,\mathtt{]}B$. Suppose that $\mathcal{M}' \models_{w_0} \mathtt{[}\,a\,\mathtt{]}B$. Let $w_0 R'_a w_1$ and $w_1 R'_b w_2$. Then $w_0 R_a w_1$ and $w_1 R_a w_2$. So $w_0 R_a w_2$, and therefore $w_0 R'_a w_2$, so $\mathcal{M}' \models_{w_2} B$. Therefore, $\mathcal{M}' \models_{w_0} \mathtt{[}\,a\,\mathtt{]}B \rightarrow \mathtt{[}\,a\,\mathtt{]}\mathtt{[}\,b\,\mathtt{]}B$.

Proof of (1.2) Let $A \in \mathrm{Tr}^a_2$. Then for some B, A is $[[a][b]B \wedge \neg Ab^a_1(B)] \rightarrow [a]B$. Suppose that $\mathcal{M}' \models_{w_0} [a][b]B$. Let $w_0 R'_a w_1$. Then $w_0 R_a w_1$, and (by Euclideanness), $w_1 R_a w_1$. Therefore, $w_1 R'_b w_1$, so $\mathcal{M}' \models_{w_1} B$. Therefore, $\mathcal{M}' \models_{w_0} [a]B$.

Proof of (2). It follows directly from Remark (i) that $\mathcal{M}'$ simultaneously satisfies T_2.

Proof of (3). Suppose $A \in T_3$; then A is $[a]B$. Let $w_0 R'_a w_1$; then $w_0 R_a w_1$. Now, $\mathcal{M} \models_{w_0} A$, so $\mathcal{M} \models_{w_1} B$. By Remark (ii), $\mathcal{M}' \models_{w_1} B$. Therefore, $\mathcal{M}' \models_{w_0} [a]B$, i.e., $\mathcal{M}' \models_{w_0} A$.

Proof of (4). Suppose $A \in T_4$; then A is $<a>B$. Now, $\mathcal{M} \models_{w_0} A$, so for some w_1, $w_0 R_a w_1$, $\mathcal{M} \models_{w_1} B$. By Remark (ii), $\mathcal{M}' \models_{w_1} B$. Therefore, $\mathcal{M}' \models_{w_0} <a>B$, i.e., $\mathcal{M}' \models_{w_0} A$.

Proof of (5). Suppose $A \in T_5$; then A is $[b]B$. Now, $\mathcal{M} \models_{w_0} A$, so $\mathcal{M} \models_{w_0} [b]B$. Let $w_0 R'_b w_1$. Then $w_0 R_b w_1$, so $\mathcal{M} \models_{w_1} B$, and by Remark (ii) we have $\mathcal{M}' \models_{w_1} B$. Therefore, $\mathcal{M}' \models_{w_0} [b]B$, i.e., $\mathcal{M}' \models [b]A$.

Proof of (6). Suppose $A \in T_6$; then A is $<b>B$. We have $\mathcal{M} \models <b>B$, so for some w_1 such that $w_0 R_b w_1$, $\mathcal{M} \models_{w_1} B$. Then $w_0 R'_b w_1$ and by Remark (ii), $\mathcal{M}' \models_{w_1} B$.

Proof of (7). Let $R_*(w)$ be the transitive closure of $\{w\}$ under R_a and R_b, and $R'_*(w)$ be the transitive closure of $\{w\}$ under R'_a and R'_b. Suppose that $A \in T_7$; then A is $[a][\mathrm{MUT}]B$. Now, $\mathcal{M} \models_{w_0} A$, so for all w_1 such that $w_0 R_a w_1$, we have $\mathcal{M} \models_{w_2} B$ for all w_2 such that $w_1 R_* w_2$. Let $w_0 R'_a w'_1$ and let $w'_1 R'_* w'_2$. Then $w_0 R_a w'_1$ and $w'_1 R_* w'_2$, so $\mathcal{M} \models_{w'_2} B$. By Remark (ii), $\mathcal{M}' \models_{w'_2} B$. Therefore, $\mathcal{M} \models_{w_0} [a][\mathrm{MUT}]B$.

By construction, $[\![Ab^a_1]\!] = [\![Ab^a_2]\!] = \emptyset$. In view of (1)-(7), $\mathcal{M}'$ is a model of T. Therefore, $\mathcal{M}'$ is a minimal model of T.

Theorem 8.2. Let T be as in Lemma 5. Then $T \mathrel{\underset{\mathrm{a}}{|\!\!\approx}} [a]A \rightarrow [a][\mathrm{MUT}]A$, for all formulas A.

Proof. In view of Lemma 5, (1) $T \mathrel{\underset{\mathrm{a}}{|\!\!\approx}} A$, for all formulas $A \in \mathrm{Tr}^a_1$ and (2) $T \mathrel{\underset{\mathrm{a}}{|\!\!\approx}} A$, for all formulas $A \in \mathrm{Normal}_{a,1}$. Then by Lemma 4, $T \mathrel{\underset{\mathrm{a}}{|\!\!\approx}} [a]A \rightarrow [a][\mathrm{MUT}]A$.

Theorem 8.3. Let T be as in Lemma 5, except that T_1 consists of all instances of the epistemic transfer axiom schemata for agent $\boldsymbol{b}$, and T_2 consists of boolean formulas that contain no occurrences of Ab^b_1 or Ab^b_2. Then $T \mathrel{\underset{\mathrm{b}}{|\!\!\approx}} [b]A \rightarrow [b][\mathrm{MUT}]A$, for all formulas A.

Proof. Just like that of Theorem 2.

Theorem 8.4. Let T be as in Lemma 5. except that T_1 consists of all instances of the epistemic transfer axiom schemata for agents $\boldsymbol{a}$ and $\boldsymbol{b}$, and T_2 consists of boolean formulas that contain no occurrences of Ab^a_1, Ab^a_2, Ab^b_1 or Ab^b_2. Then $T \mathrel{\underset{\mathrm{a,b}}{|\!\!\approx}} [[a]A \wedge [b]A] \rightarrow [\mathrm{MUT}]A$, for all formulas A.

Proof. Combine the constructions used in the proofs of Theorem 2 and Theorem 3, and use Lemma 3 to establish the result.

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 reasoning that governs belief transfer will be blocked.

The example I'll develop in this section resembles the one at the beginning of Clark & Marshall [5] which shows that the iteration complexity for agent knowlege can reach fairly high finite levels. That example, though, deals simply with agent beliefs. As I explained in Section 2.2.2, I feel 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, there are two agents, $\boldsymbol{a}$ and $\boldsymbol{b}$ in the following example. 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 $\mathrm{ROOT} \preceq \mathrm{PUB}$, but $\mathrm{PUB} \not\preceq \mathrm{ROOT}$.

The following rudimentary theory of email communication consists of three parts: (A) protocols for updating the contents of $[a, \mathrm{PUB}]$, (B) a theory of exceptions to the protocols in (1), and (C) the transfer axiom schemata (Tr^a_1) and (Tr^a_2). To keep things simple, the formalization ignores temporal considerations. The following axiom schemata use quantification over individuals (but not over propositions).

(A) **Protocols for updating** $[a, \mathrm{PUB}]$

(A.1) $\forall m[[\mathit{Send}(a,b,m) \wedge \mathit{Incontents}(m,A)] \rightarrow [a, \mathrm{PUB}]A]$

If $\boldsymbol{a}$ sends a message to $\boldsymbol{b}$ that says p (where, as before, A expresses the proposition p, then $\boldsymbol{a}$ adds p to $[a, \mathrm{PUB}]$.

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

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

Notes on the formalization. Neither (A.1) nor (A.2) is a default. The nonmonotonic aspects of the reasoning about belief are handled by the epistemic transfer rules. This division of labor seems simpler on the whole; it has the defect that the modality $[a, \mathrm{PUB}]$ no longer represents $\boldsymbol{a}$'s view of $\boldsymbol{b}$'s beliefs. Since other modalities, such as $[a, \mathrm{PUB}][b, \mathrm{PUB}]$ and $[\mathrm{MUT}, \mathrm{PUB}]$ are available, this is not much of a disadvantage. In order to prove that mutuality is secured by default, we may wish to restrict the values of 'A' in these schemata to boolean formulas that do not contain abnormality predicates.

(B) **The abnormality theory.**

(B.1) $\forall m_1 \forall m_2[\mathit{Send}(a,b,m_1) \wedge \mathit{Incontents}(m_1,A) \wedge \mathit{Read}(a,m_2)$
$\wedge\, \mathit{Not\text{-}Delivered}(m_2,m_1)] \rightarrow \mathit{Ab}^a_1(A)]$

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

(B.2) $\mathit{Ab}^a_1(A) \rightarrow \mathit{Ab}^a_1([a, \mathrm{PUB}]A)$

If a proposition is abnormal, so is the proposition that $\boldsymbol{a}$ publically believes it. (This is needed for technical reasons; without it, $[a, \textsc{pub}][\textsc{mut}]MB$ could be inferred in Case 2, below.) The next schema is similar.

(B.3) $Ab_2^a(A) \rightarrow Ab_2^a([a, \textsc{pub}]A)$

(C) **The transfer axiom schemata.**

(Tr$_1^a$) $[\neg Ab_1^a(A) \wedge [a, \textsc{pub}]A] \rightarrow [a, \textsc{pub}][b, \textsc{pub}]A]$

(Tr$_2^a$) $[\neg Ab_2^a(A) \wedge [a, \textsc{pub}][b, \textsc{pub}]A] \rightarrow [a, \textsc{pub}]A]$

I will present four cases, of increasing complexity.

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 $[a, \textsc{pub}]$.

Here, we want $[a, \textsc{pub}][\textsc{mut}]MB$ to be a circumscriptive consequence of the theory—we want the theory to imply that Ann (publically) believes that the content of the message is (publically) mutually believed. Theorem 2 does not suffice to prove this, since now a part of the theory deals with abnormality. A construction similar to the one used in the proof of Lemma 5 provides the desired result. The presence of an abnormality theory provides one further complication—Definition 21 needs to be amended so that the predicates on which abnormalities depend, *Send*, *Incontents*, *Read*, *Not-Delivered*, and *Ack*, are varied in circumscribing for Ann's beliefs about Bob's beliefs. Allowing these predicates to vary arbitrarily is somewhat implausible, and it may well be necessary to appeal to a more powerful circumscriptive technique, such as pointwise circumscription (see Lifschitz [15]). The possible need for such techniques is one side effect of the complexity of these examples.

The formalization:

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

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

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

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 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 despite the mailer daemon message, Bob has received a copy of M_2.

The reasoning:

As in Cases 1 and 2, Ann adds MB to the beliefs of the subagent representing *prima facie* beliefs shared with Bob. As in Case 2, receipt of the mailer daemon's message precipitates an abnormality of the form $Ab_1^a(MB)$, which blocks the usual chain of inference to $[\,b, \textsc{pub}\,]MB$. But Bob's acknowledgement provides direct evidence for $[\,b, \textsc{pub}\,]MB$; the protocol axiom (A.2), permits $[\,a, \textsc{pub}\,][\,b, \textsc{pub}\,]MB$ to be inferred without using (Tr_1^a).

Using techniques from the proof Lemma 4, I believe that we can show that under these circumstances, $[\,a, \textsc{pub}\,][\,\textsc{mut}, \textsc{pub}\,]MB$ can be inferred, provided that we also have $\neg Ab_1^a(MB)$ and $\neg Ab([\,\alpha\,]MB)$, for any string α containing at least one occurrence of $\boldsymbol{b}$. To obtain Ann's belief of the message's mutuality as a circumscriptive consequence, we would need to show that in any minimal model of the theory, these abnormalities are indeed empty. I believe this can be done, but I have not checked all the details.

Case 4. 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: This case would require the addition of an abnormality theory for Ab_1^a; I have not provided such a theory. The desired conclusions here would include:

1. Ann (publically) believes p.
2. Ann (publically) believes that Bob (publically) believes p.

But they would not include

3. Ann (publically) believes that Bob (publically) believes that Ann (publically) believes p.

The reasoning here is about as complicated as common sense reasoning about attitudes can get. The cases in this example are meant to illustrate a sequence of increasingly complex formalization problems; this last one is best thought of as a challenge. I confess that I have not yet thought through the formalization issues for this case.

10. Conclusion

Previous attempts to explain interagent reasoning about attitudes do not provide plausible formalizations of the reasoning that underlies mutuality in cases that seem to require it, or provide logical resources 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 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 exhibit 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 do not claim to have done that here, and in fact the cases considered in Section 9 put a certain degree of pressure on the circumscriptive approach that is developed in this paper. Obviously, more work needs to be done on the logical foundations, to develop an approach to the nonmonotonic reasoning that is not intolerably complex and is flexible enough to handle cases like those of Section 9. I myself am happy to be eclectic about the choice of nonmonotonic formalisms, and I believe that alternatives to the circumscriptive approach need to be explored.

I hope that at least I have made a plausible case for the promise of the general approach that is developed in this paper, and in particular that I have convinced you that immediacy theories of how mutuality is secured do not do justice to the relevant reasoning.

Of course, it is highly desirable not only to deploy nonmonotonic formalisms in applications such as this, but to show how the reasoning can be efficiently implemented in special cases. I have not addressed that question in this paper, but I hope to do this in subsequent work.

References

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

[2] 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.

[3] 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.

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

[5] 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.

[6] Herbert H. Clark and Michael Schober. Understanding by addressees and overhearers. *Cognitive Psychology*, 24:259–294, 1989. Republished in [4].

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

[8] Kit Fine. Propositional quantifiers in modal logic. *Theoria*, 36:336–346, 1970.

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

[10] Joseph Y. Halpern and Gerhard Lakemeyer. Multi-agent only knowing. In Yoav Shoham, editor, *Theoretical Aspects of Rationality and Knowledge: Proceedings of the Sixth Conference (TARK 1996)*, pages 251–265. Morgan Kaufmann, San Francisco, 1996.

[11] Philip Kremer. On the complexity of propositional quantification in intuitionistic logic. *The Journal of Symbolic Logic*, 62(2):529–544, 1997.

[12] 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.

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

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

[15] Vladimir Lifschitz. Pointwise circumscription. In Tom Kehler and Stan Rosenschein, editors, *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 406–410, Los Altos, California, 1986. American Association for Artificial Intelligence, Morgan Kaufmann.

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

[17] Ana Maria Monteiro and Jacques Wainer. Preferential multi-agent nonmonotonic logics. In Luigia Carlucci Aiello, Jon Doyle, and Stuart Shapiro, editors, *KR'96: Principles of Knowledge Representation and Reasoning*, pages 446–452. Morgan Kaufmann, San Francisco, California, 1996.

[18] Leora Morgenstern. A theory of multiple agent nonmonotonic reasoning. In Thomas Dietterich and William Swartout, editors, *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 538–544, Menlo Park, CA, 1990. American Association for Artificial Intelligence, AAAI Press.

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

[20] Rohit Parikh. Monotonic and nonmonotonic logics of knowledge. *Fundamenta Informaticae*, 15(3–4):255–274, 1991.

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

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

[23] Richmond H. Thomason. Propagating epistemic coordination through mutual defaults I. In Rohit Parikh, editor, *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the Third Conference (TARK 1990)*, pages 29–39, Los Altos, California, 1990. Morgan Kaufmann.

[24] 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.

[25] Jacques Wainer. Epistemic extension of propositional preference logics. In Ruzena Bajcsy, editor, *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 382–387, San Mateo, California, 1993. Morgan Kaufmann.