Intra-Agent Modality and Nonmonotonic Epistemic Logic

Richmond H. Thomason
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260
U.S.A.

thomason@isp.pitt.edu
http://www.pitt.edu/~thomason/thomason.html

1. Background and Motivation

It is plausible to think that simulation is perhaps the most important reasoning tool that we have for user modeling. This is behind what we mean when we say that a superlative fisherman can "think like a fish." The fisherman decides where the fish must be by imagining where *he* would be in this river if he were a fish.

Whether or not this idea is sound for fish and fisherman,[1] it certainly applies with a great deal of force to people reasoning about one another's attitudes, preferences, emotions, and choices. A friend tells me a story about problems she's been having with her car. She seems quite calm, but I say "You must be upset," reasoning that if this happened to me, *I* would be upset. I go on, saying "You must realize your mechanic is lying to you" because her description of the problem indicates she knows as much about cars and mechanics as I do, and knowing what she has told me, *I* would infer that her mechanic is lying. This sort of *other-modeling* is the reasoning that makes the "golden rule" golden. What would be the moral point of doing unto others as you would have others do unto you if imagining what we ourselves would want were an unreliable way to gauge what others want?

In a number of psycholinguistic investigations, Herbert Clark has demonstrated many ways in which conversation is informed by common ground. The following account of how conversants construct common ground is taken from Clark & Schober [4, pp. 257–158]. (Page numbers from the version in *Arenas of Language Use.*)

> 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.

[1] This is a by now classic topic in contemporary philosophy of mind and consciousness; see Nagel [18], and, for instance, Baars [1].

> 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 ...
>
> ... People must keep track of communal and personal common ground in different ways. For communal common ground, they need encyclopedias for each of the communities they belong to. Once Alan and Barbara establish the mutual belief that they are both physicians, they can immediately add their physician encyclopedias to their common ground.

This account of common ground is compelling and plausible.[2] The purpose of this paper is to develop a logical theory of this sort of reasoning. I believe that the most important part of developing such a theory is to begin with a model of single-agent attitudes that makes this sort of other-modeling possible.

Take as an example the case of modeling beliefs. If other agents' beliefs were exactly like ours, we could form conclusions about them by imitation, by simply consulting our own beliefs. As it is, however, the beliefs of others differ from our own, so if we are to use imitation for other-modeling, we must somehow be able to adjust our beliefs. But there are independent reasons for thinking we have this ability. Thomason [20] examines ways in which many human attitudes, including belief, are sensitive to context—where context includes not only purely epistemological factors like available evidence, but also matters like the risk of acting on a supposition, the time available for deliberation, and factors affecting the power of wishful thinking. In light of these considerations, it seems better to begin with a flexible sort of supposition, which is affected by various contextual factors. Belief can then be treated as a form of supposition on which an agent is willing to plan and act.

So an agent a_1 that is capable of modeling the beliefs of another agent a_2 should, firstly, be capable of simulating a variety of belief operators, $\Box_1$, $\Box_2, \ldots$.[3] Secondly, using information about a_2, it should have a way to select one of these internal belief attitudes $\Box_i$ for representing a_2's beliefs (or, more accurately, part of a_2's beliefs). Then a_1's model of a_2's beliefs is simply $\Box_i$. To show that a_2 believes A, a_1 establishes that *it itself* believes A, modulo $\Box_i$, using whatever method it has available for verifying that it has a belief. To show that a_2 does not believe A, a_1 establishes that it itself does not believe A, modulo $\Box_i$, using whatever method it has available for establishing that it lacks a belief.

As a first approximation, this amounts to saying that belief-modeling is guided by the following axiom scheme, for a suitably chosen modality $\Box_i$:[4]

[2] The metaphor of the encyclopedia that is mentioned in Clark and Schober's account is, however, a little misleading. It would be more accurate, I think, to take an object-oriented approach to the indexing of beliefs. The idea is that we maintain a hierarchy of communities to which beliefs can be indexed; beliefs with general indices (e.g., beliefs indexed to US citizens) are available at more specific indices (e.g., at the index for Californians). When a belief is acquired, it is assigned one or more indices, which provide information about what communities can be expected to have the belief. For instance, the information that is acquired in general courses in medical school would be assigned a PHYSICIAN index.

[3] In this paper, I will use $\Box$ as a generic modal operator, including non-alethic modalities like belief and supposition.

[4] This is an oversimplification. In fact we can't hope to model another's beliefs using (1.1) if there is

(1.1) $\Box_{a_1}[\Box_{a_2}A \leftrightarrow \Box_i A]$.

In this paper, I will do three things: (i) I will explore the consequences for single-agent epistemic logic of assuming that agents are capable of this sort of reasoning, (ii) I will extend the single-agent logic to the monotonic multi-agent case, and (iii) I will indicate (very briefly) how the reasoning can be situated in a nonmonotonic framework. The technical apparatus that emerges is similar to the formalizations of contextual reasoning inspired by McCarthy [13].[5]

2. Modeling the Multiplicity of Single-Agent Beliefs

2.1. A Logical Model

In a variety of reasoning tasks, it has proved to be important to keep track not only of bare claims, but of the *support* for these claims.[6] Humans must find it useful to do much the same thing (probably, because it makes belief revision and learning much easier). The question "How do you know that?" makes sense with respect to a large number of claims, and in a remarkably large number of cases we are able to answer such questions.

In compiling belief into a single modality, the standard representation of belief in epistemic logic loses information about the antecedents of beliefs. A simple way to restore the missing information, which preserves the framework of modal logic, is to replace a single unanalyzed belief modality with a family of modalities corresponding to different sources of information. For present purposes, we do not need to distinguish the source of a belief from a collection of basic reasons.

If we think of it axiomatically, the idea is that set of all beliefs is like an axiomatized theory that is modularized into subtheories. This organization does not affect the total quantity of derived information, but (if the modularization is properly designed) may make the theory easier to understand and maintain. There is no restriction concerning the content of the various subtheories; they may deal with specific topics (e.g., one subtheory may deal with arithmetic, another with geometry) or with interrelationships between topics (e.g. a subtheory may present the Cartesian rules for modeling the Euclidean plane).

This division into subtheories can be marked in the object language by means of modalities. The language is just like the multiagent modal logics that have become current in modeling communications protocols and games.[7] Each subtheory is assigned an index; and a theory of inter-index relations determines information relations among indices. In multiagent modal logic, the distributed systems application is primary; message-passing is the chief epistemic relation between indices, reasoning about other agents is crucial, and mutual attitudes are important. In intra-agent modal logic, at least as I want to explore it, forms

any uncertainty about these beliefs. Moreover, (1.1) is asymmetric; it treats beliefs of the modeler a_1 as modular, and the beliefs of the modeled a_2 as monolithic. These defects are related: for a better treatment, see Example 1, below.

[5] This should not surprise readers familiar with the recent AI literature on context. I have to confess, however, that the extent of the parallel only struck me after I had thought for some time about the problem of belief modeling.

[6] See, for instance, DeKleer [5] and Mitchell [14].

[7] See Fagin et al. [7].

of *access* are the primary epistemic relations between indices, the purpose is to access information from other agents[8] rather than to reason about them, and hierarchical relations between modalities are important.

At this point, I will introduce new notation for the indexed epistemic operators: $[i]$ for "i believes" and $<i>$ for "for all i believes".[9]

Call the indices of an intra-agent model *subagents*; They are in some ways analogous to the agents of multi-agent modal logic, but we have to bear in mind that they merely represent convenient modularizations of an agent's beliefs. 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. In the applications that I have in mind, there is no interaction other than access between subagents. 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 $[i][j]A$, but linguistic restrictions of this kind are in general less satisfactory than a semantic treatment. We might make statements about j's beliefs to be neither true nor false for i. But truth-value gaps introduce more complications than they are worth. Here, I will assume that $[i][j]A$ is false if i can't access j.

These ideas lead to the following definition.

Definition 2.1. Intra-Agent Modal Languages.

An intra-agent propositional language $\mathcal{L}(\mathcal{I}, \preceq, \mathcal{P})$ 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 2.2. Intra-Agent Modal Formulas.

Where $i \in \mathcal{I}$, 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. I.e., for all $i \in \mathcal{I}$, if $A \in \textsc{Formulas}(\mathcal{P}, \mathcal{I})$, then $[i]A \in \textsc{Formulas}(\mathcal{P}, \mathcal{I})$.

Definition 2.3. Intra-Agent Modal Frames.

An *intra-agent frame* $\mathcal{F}(W, \mathcal{I}, R)$ consists of (1) a nonempty set W of possible worlds, (2) the reflexive, transitive ordering $\preceq$ over $\mathcal{I}$, and (3) a relation R_i over W for each $i \in \mathcal{I}$.

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 $i \preceq j$.

[8] Absorption is a metaphor for methods of information transfer that depend on the agent's architecture. For a very general modular approach to agent modeling, see Doyle [6].

[9] I believe that what follows is general with respect to the distinction between knowledge and belief, and in general does not depend on what conditions one wishes to place on the single-agent accessibility relations. The logics discussed below are all variations on the modal logic **DS5**, but the choice of this modality is largely for concreteness and for illustrative purposes.

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]. A fundamental assumption of the approach that I am taking here is that we can model the intra-agent modularization of belief with the same basic logic that is used for multi-agent epistemic logic, together with additional constraints that are appropriate for the intra-agent case. I believe that Subagent Monotonicity and Subagent Coherence provide the needed additional constraints.

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. This means that, in particular, i knows everything that j knows. Therefore, every world that is i-entertainable is j-entertainable; this is Subagent Monotonicity. Furthermore, i must represent j's beliefs as true; a world that is not j-entertainable relative to itself is not i-entertainable relative to any world. This is Subagent Coherence. Although I am interested in models that satisfy all of these constraints, I have tried to axiomatize the logic in a way that separates the constraints. Axioms 0–5 below, and Rules 0–1, hold in all models, regardless of constraints. Axiom 6 corresponds to Transitivity, Axiom 7 to Euclideanness, Axiom 8 to Seriality, Axiom 9 to Subagent Monotonicity, and Axiom 10 to Subagent Coherence.

Definition 2.4. Intra-Agent Modal Models.
An *intra-agent modal model* $\mathcal{M} = \langle W, R, V\rangle$ of an intra-agent modal language $\mathcal{L}(\mathcal{I}, \preceq, \mathcal{P})$ consists of an intra-agent frame $\mathcal{F}(W, \mathcal{I}, R)$ and a valuation V, where V is a function from $\mathcal{I} \times \mathcal{P}$ to $\{T, F\}$.

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 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 2.5. Satisfaction in an Intra-Agent Modal System.
Satisfaction is standard for boolean connectives, and $\mathcal{M} \models_{i,w} \texttt{[}j\texttt{]}A$ iff $i \preceq j$ and for all $w \in R_jw'$, $\mathcal{M} \models_{j,w'} A$.

This logic may seem too peculiar to fit into the spectrum of known modal logics. (Note, for example, that no formula of the form $\texttt{[}i\texttt{]}A$ is valid.) But in fact, the single-subagent case, where $\mathcal{I} = \{i\}$, is a deontic **S5**-like version of the non-normal logic **E2** that is formulated in Lemmon [12] and proved complete in Kripke [11]. 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. In particular, formulas of the sort $\texttt{[}j\texttt{]}\top$ hold at $\langle i, w\rangle$ if and only if $i \preceq j$, and so can be used to keep track of subagent accessibility relations.

2.2. Incompactness

The full logic has another peculiarity, that may at first be surprising. It is incompact.

Theorem 2.1. Let $\mathcal{I} = \{i_0, i_1, \ldots\}$. Suppose that $\preceq$ is nontrivial, in the sense that for all $\mathcal{J} \subseteq \mathcal{I}$ there is a k such that $k \not\preceq j$ for each $j \in \mathcal{J}$. (Note that $\mathcal{I}$ is nontrivial if there is any sequence $j_0 \preceq j_1 \preceq j_2 \preceq \ldots$ with $j_i \in \mathcal{I}$.) Then $\mathcal{L}(\mathcal{I}, \preceq, \mathcal{P})$ is incompact.

Proof. Let $\Gamma_n = \{\neg[i_m]\top : m \leq n\}$, and let $\Gamma = \bigcup\{\Gamma_n : n \in \omega\}$. Now, Γ is not satisfiable, for if $\mathcal{M}$ satisfies Γ then for some $i \in \mathcal{I}$, we would have $\mathcal{M} \models_{i,w} \neg[j]\top$ for all $j \in \mathcal{I}$. This is impossible, since clearly $\mathcal{M} \models_{i,w} [i]\top$. On the other hand, every Γ_n is satisfiable, simply by choosing k so that $k \neg \preceq i_m$ for all $m \leq n$; then for any $\mathcal{M}$, $\mathcal{M} \models_{k,w} A$ for all $A \in \Gamma_n$.

The incompactness result provides a good reason for assuming that $\mathcal{I}$ is finite. From now on, I will make this assumption.

2.3. Axiomatization

R0. $$\frac{A \quad A \to B}{B}$$ **R1.** $$\frac{A \to B}{[i]A \to [i]B}$$

Ax0. Any substitution instance of a boolean tautology.
Ax1. $[i][A \to B] \to [[i]A \to [i]B]$
Ax2. $[i]\top \to [j]\top$ if $i \leq j$
Ax3. $[i]\top \to \neg[j]\top$ if $i \not\preceq j$
Ax4. $[i]\top \to [i][i]\top$
Ax5. $\bigvee\{[i]\top \wedge \bigwedge\{\neg[j]\top : i \not\preceq j\} : i \in \mathcal{I}\}$
Ax6. $[i]A \to [i][i]A$
Ax7. $\neg[i]A \to [i]\neg[i]A$
Ax8. $\langle i \rangle \top$
Ax9. $[[j]A \wedge [i]\top] \to [i]A$ if $i \preceq j$
Ax10. $[i]\top \to [i][[j]A \to A]$

2.4. Soundness and Completeness

I have a detailed proof, in handwritten form, of soundness and completeness of these axioms with respect to the intended semantics. I will make it available to anyone who is interested. (My email address and web page are listed at the beginning of this paper; check my web page to see if the proof has been formatted and put online.) The completeness proof resembles familiar Henkin-style proofs for modal logics in most respects, but has a few unusual features.

2.5. An Example: Public and Private Beliefs

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.

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:

$$\preceq = \{\langle \text{MIN}, \text{MIN}\rangle, \langle \text{MIN}, \text{PUB}\rangle, \langle \text{MIN}, \text{NPUB}\rangle, \langle \text{PUB}, \text{PUB}\rangle, \langle \text{NPUB}, \text{NPUB}\rangle\}.$$

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.

$R_{\text{NPUB}}(w, w')$ iff $w, w' \in \{w_1, w_3\}$.
$R_{\text{PUB}}(w, w')$ 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" $\langle \text{MIN}, w\rangle$ 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. $[\text{PUB}]p_1$, $\neg[\text{NPUB}]p_2$
4. $[\text{MIN}][p_1 \wedge p_2]$, $[\text{MIN}][\text{NPUB}]p_2$, $[\text{MIN}]\neg[\text{NPUB}]p_1$, $[\text{MIN}][\text{PUB}]p_1$, $[\text{MIN}]\neg[\text{NPUB}]p_2$

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.

3. Modal Logic for Multi-Agent Beliefs

We now want to imagine a community of agents. Each of them 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].

3.1. Extending the Language to the Multi-Agent Case

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

$$[a,i][P \to [b,j][Q \to [a,i]R]],$$

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. (This assumption simplifies things, but is not necessary.)

The notation may appear to assume that each subagent knows about each other agent's modular structure, but in fact the assumption is only 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 3.6. Inter-Agent Modal Languages.

An inter-agent propositional language $\mathcal{L}(\mathcal{P},\mathcal{I},\mathcal{A},\preceq)$ 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 $\mathcal{A}$); and by a function $\preceq$ which for each $a \in \mathcal{A}$ provides a reflexive, transitive ordering on $\mathcal{A}$. We assume that if $a \neq b$, then $\mathcal{I}_a$ and $\mathcal{I}_b$ are disjoint. Where $i \in \mathcal{I}$ and $a \in \mathcal{A}$, the set $\textsc{Formulas}(\mathcal{P},\mathcal{I},\mathcal{A})$ is the smallest set that (1) contains $\mathcal{P}$, (2) is closed under boolean connectives, and (3) is closed under i,a-necessitation. I.e., for all $a \in \mathcal{A}$, if $A \in \textsc{Formulas}(\mathcal{P},\mathcal{I},\mathcal{A})$, then $[a,i]A \in \textsc{Formulas}(\mathcal{P},\mathcal{I})$, where $i \in \mathcal{I}_a$.

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

Definition 3.7. Inter-Agent Modal Frames.

An *inter-agent frame* $\mathcal{F}(W,R)$ for $\mathcal{L}(\mathcal{P},\mathcal{I},\mathcal{A},\preceq)$ consists of (1) a nonempty set W of possible worlds, (2) the reflexive, transitive ordering $\preceq$ over $\mathcal{I}$, and (3) a relation R_i over W for each $i \in \mathcal{I}_a$, where $a \in \mathcal{A}$.

The meaning of the subagent accessibility relation is slightly different in the multi-agent logic. In case i and j are subagents of the same agent, then $i \preceq j$ means that i accesses the beliefs of j, as in the pure intra-agent case. In case i and j are subagents of different agents, then $i \preceq j$ means that subagent i of agent a is able to model the beliefs of subagent j of agent b, where $a \neq b$. This means that i has opinions about j's beliefs, not that j represents a component of i's beliefs.

Definition 3.8. Inter-Agent Modal Models.

The satisfaction relation $\mathcal{M} \models_{i,w} A$ is relativized to subagents and to worlds, as before; satisfaction conditions are standard for boolean connectives, and $\mathcal{M} \models_{i,w} [a,j]A$ iff $i \preceq j$ and $\mathcal{M} \models_{j,w'} A$ for all w' such that wR_jw'.

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

As before, we are primarily interested in models that are Transitive, Euclidean, Serial, and Subagent Monotonic and Coherent. The latter two conditions appear as follows in the general setting.

Subagent Monotonicity. $R_i \subseteq R_j$ if for some a, $i, j \in \mathcal{I}_a$ and $i \preceq j$.

Subagent Coherence. If wR_iw', for some a, $i, j \in \mathcal{I}_a$, and $i \preceq j$ then $w'R_jw'$.

The axiomatization of this logic is a straightforward extension of the pure intra-agent case; details will not be presented here, but are available on request.

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

3.2. Adding Mutual Belief

We can add a mutual belief operator $[c]$ in the usual way. Here, I will consider a mutual belief operator only for the entire group $\{i_a : a \in \mathcal{A}\}$ of agents. This is a simplification; in the general case, special subagents capable of modeling one another will generate special-purpose mutual belief operators. (Recall the quotation from Clark & Schober [4, pp. 257–158] with which we began; to model these ideas we would create subagents keeping track of beliefs common to the specific cultural communities; each of these subagents would create a mutual belief operator.)

Definition 3.9. Inter-Agent Modal Systems with Mutual Belief.
An inter-agent propositional language $\mathcal{L}(\mathcal{P}, \mathcal{I}, \mathcal{A})$ with mutual belief is a general inter-agent propositional language with an a modal operator $[\text{MUT}]$. The satisfaction condition for $[\text{MUT}]$ is as follows:

$\mathcal{M} \models_{i,w} [\text{MUT}]A$ iff $\mathcal{M} \models_{i,w'} A$ for all w' such that wR_cw', where R_c is the transitive closure of the set of relations $\{R_{i_a} : a \in \mathcal{A}\}$.

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. [7]; our operators $[i_a]$ correspond to the agent belief operators K_i of that system.[10]

4. Nonmonotonic Reasoning about Beliefs

There is only space in the present paper for a brief sketch of the nonmonotonic case.

Several frameworks have been proposed for formalizing nonmonotonic reasoning about beliefs: autoepistemic logic, Morgenstern [17]; circumscription in a higher-order modal logic,

[10] Note that here, however, we are using indices like 'i' and 'j' for subagents, and indices like 'a' and 'b' for agents.

Thomason [21]; default logic (or something like it), Parikh [19]; only-knowing, Halpern & Lakemeyer [9]; and preferential models, Wainer [22] Monteiro & Wainer [16].

I will present a circumscriptive approach, mainly because I believe that circumscription is a useful and straightforward development tool. 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 circumscriptive approach to nonmonotonic reasoning appeals to a circumscription operation on finitely axiomatized theories, which takes the original theory into one in which certain terms are minimized. If we consider theories that involve *epistemic transfer rules* like (1.1), the need for finitely axiomatized theories requires a logic with quantifiers over agents and over propositions. This can be accomplished, as in Thomason [21], by embedding the logic in a modal type theory along the lines of Montague [15] and Gallin [8]. Even if we restrict the logic to one with only quantifiers over propositions and agents, we will have an unaxiomatizable logic; see, for instance, Kremer [10]. I believe that some such complexity is an inevitable consequence of a general approach to the phenomenon of knowledge transfer.[11] Hopefully, however, tractable special cases will emerge.

A simple example of a transfer rule is the principle stating that agents believe that normally, if a "public" subagent believes something, then all public subagents share this belief. This rule can be formulated as follows.[12]

$$(4.1)\ \ \forall a \forall b \forall p [\, i(a)\,][[[\, i(a), \textsc{pub}\,]p \wedge \neg Ab(p,a,b)] \rightarrow [b, \textsc{pub}]p].$$

This axiom uses the abnormality predicate Ab to represent the abnormality of a proposition for agents a and b. In specific cases, we can provide special axioms about abnormality; for instance, axioms to the effect that $Ab(p, x, y)$ holds for proposition p and agents a and b in case a has heard b express doubt as to p.

We need separate axioms to enable agents to reach conclusions about other agents' ignorance. Here is one possibility.

$$(4.2)\ \ \forall a \forall b \forall p [\, i(a)\,][[[\, i(a), \textsc{npub}\,]p \wedge \neg [\, i(a), \textsc{pub}\,]p \neg Ab'(p,a,b)] \rightarrow [b, \textsc{pub}]p].$$

When I learn something by a private channel—for instance when I observe a fleeting event under private circumstances—I use the NPUB subagent to record this. I then normally assume that any information that the private NPUB subagent is aware of and that the public PUB subagent is *not* aware of is not believed by others. This axiom will work fairly well, as long as I am careful to add information to the public module whenever I publish it; for instance, if I tell a friend about it. Once information is made public in this fashion, it becomes diffcult to write plausible axioms about who can be expected to be aware of it. This seems to correspond to the actual facts; it is not easy in general to reason reliably about who might be aware of semi-public information.

Using versions of these circumscriptive axioms, it is possible to show that, under highly idealized circumstances, communication can lead to coordinated belief; subagents of different agents will have the same beliefs, and these beliefs will be mutual. There is a result of this

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

[12] Here, $i(a)$ is an object-language representation of i_a.

kind, in fact, in Thomason [21]. But there is no free lunch here: we have to assume that different agents are able to initialize the conversation by creating subagents with the same beliefs, that they have the same nonmonotonic rules for processing conversation, and that there are no "mishearings." These idealizations are, of course, too extreme to provide a realistic model of actual communication.

5. Conclusion

There is nothing new about nonmonotonic approaches to reasoning about knowledge, and nothing very new in the nonmonotonic aspects of the framework sketched above in Section 4. However, previous attempts to model nonmonotonic 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 the sort of reasoning cited in Clark & Schober [4]. Unless I have missed something, the literature contains no very promising way to provide specific, formal reasoning mechanisms for obtaining mutuality.

In Barwise [2], the suggestion is made that mutuality somehow arises out of a shared situation, and the way in which this happens is left as a mystery. This suggestion is of little or no value, since shared situations do not in general lead to mutuality—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 mechanisms that allows agents to obtain it from experience, in some cases, from knowledge 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 sketched in this paper.

Of course, it is highly desirable not only to deploy these formalisms, but to show how the reasoning can be efficiently implemented in special cases. Modal theorem proving would be a more or less direct way to implement many of the mechanisms that are discussed here, but more generally any form of declarative reasoning in which knowledge is partitioned along modal lines could provide an approximate implementation. At this point, however, I have to confess that I have given very little thought to reasoning issues.

Bibliography

[1] Bernard J. Baars. A thoroughly empirical approach to consciousness. *Psyche*, 6, 1994.

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

[4] Herbert H. Clark and Michael Schober. Understanding by addressees and overhearers. *Cognitive Psychology*, 24:259–294, 1989. Republished in [3].

[5] Johan de Kleer. An assumption-based TMS. *Artificial Intelligence*, 26(1):127–162, 1985.

[6] Jon Doyle. A society of mind—multiple perspectives, reasoned assumptions, and virtual copies. In Barbara Hayes-Roth and Richard Korf, editors, *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 309–313, Menlo Park, CA, 1994. American Association for Artificial Intelligence, AAAI Press.

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

[8] Daniel Gallin. *Intensional and Higher-Order Logic.* North-Holland Publishing Company, Amsterdam, 1975.

[9] 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.

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

[11] 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.

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

[13] John McCarthy. Notes on formalizing contexts. In Tom Kehler and Stan Rosenschein, editors, *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 555–560, Los Altos, California, 1986. American Association for Artificial Intelligence, Morgan Kaufmann.

[14] Tom M. Mitchell. *Version Space: An Approach to Concept Learning.* Ph.d. dissertation, Computer Science Department, Stanford University, Stanford, California, 1986.

[15] Richard Montague. Pragmatics and intensional logic. *Synthèse*, 22:68–94, 1970. Reprinted in *Formal Philosophy*, by Richard Montague, Yale University Press, New Haven, CT, 1974, pp. 119–147.

[16] 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.

[17] 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.

[18] Thomas Nagel. What is it like to be a bat. *The Philosophical Review*, 83:435–450, 1974.

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

[20] Richmond H. Thomason. The context-sensitivity of belief and desire. In Michael P. Georgeff and Amy Lansky, editors, *Reasoning About Actions and Plans*, pages 341–360. Morgan Kaufmann, Los Altos, California, 1986.

[21] 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.

[22] 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.