A Formal Theory of Multiple Agent Nonmonotonic Reasoning

Leora Morgenstern
leora@ibm.com
IBM T.J. Watson Research
P.O. Box 704, Mail Stop H1N08
Yorktown Heights, N.Y. 10598
(914)784-7151

Abstract

This paper presents a formal theory of multiple agent non-monotonic reasoning. We introduce the subject of multiple agent non-monotonic reasoning for inquiry and motivate the field in terms of its applications for commonsense reasoning. We extend Moore's [1985] Autoepistemic Logic to the multiple agent case, and show that the resulting logic is too weak for most applications of commonsense reasoning. We then suggest some possible sets of principles for a logic of multiple-agent non-monotonic reasoning, based on the concept of an agent's *arrogance* towards his knowledge of another agent's ignorance. While the principles of arrogance are in general too strong, we demonstrate that restricted versions of these principles can work quite well for commonsense reasoning. In particular, we show that a restricted form of the principle of arrogance yields results that are equivalent to EMAT [Morgenstern, 1989], a non-monotonic logic which was designed to reason about temporal projection in epistemic contexts.

1. Introduction and Motivation

Past research in non-monotonic reasoning (cf. [Ginsberg, 1987]) has concentrated almost exclusively on the single-agent case. [1] The focus has been either on how a single agent uses default rules of reasoning ([Reiter, 1980], [McDermott and Doyle, 1980]), or how an agent reasons about his own beliefs [Moore, 1985]. In fact, much practical non-monotonic reasoning involves reasoning about multiple agents. An agent who wishes to predict what another agent believes must reason about how the second agent's reasoning is non-monotonic. If I tell Susan that I have just bought a bird, I should be able to predict that Susan will believe that the bird flies. In order to do that, I will have to understand how Susan reasons with default rules and to know the default rules that Susan uses.

The ability to reason about how other agents reason non-monotonically is particularly crucial for multi-agent planning domains. An agent in a multi-agent domain who constructs any complex plan will have to reason about the ways in which other agents' actions interact with his plan. That is, he must be able to predict how other agents will act. This, in turn, entails having some set of beliefs about other agents' beliefs and goals.

For example, suppose that I plan to meet Carol at a restaurant five blocks from my home at 8 P.M. Carol lives 20 miles away, and I know that she assumes that it typically takes 40 minutes to travel that particular stretch of road. I can thus reason that Carol leaves her home at approximately 7:20 P.M. If at 7:40, I turn on the radio and discover that there is a major traffic jam on the route that Carol takes, I can conclude that Carol will be quite late. Most likely, I will not go down to the restaurant at 8, but will postpone going down until sometime later. In order to engage in this sort of reasoning, it is necessary for me to reason about Carol's default beliefs regarding the time it takes her to travel.

Other more complex examples involve the frame problem and temporal projection. Agents who plan must reason about what stays the same when actions are performed. Typically, agents perform temporal projection by doing some sort of non-monotonic reasoning. If I count on someone else to perform an action in my plan, I must be able to reason about how he performs temporal projection. Thus, I must be able to reason about how he reasons non-monotonically. A specific example of this sort of reasoning is given in Section 3.

Another rich source of examples comes from speech acts theory. Gricean theory [Grice, 1957] is best modelled in terms of mutual default assumptions about the other agents' beliefs regarding the conventions that they both hold. Russell [1987] has argued that communication is enabled by a *mutual absence of doubt* on the part of the agents that they have different conven-

[1] A notable exception is the work of [Perrault, 1987] and [Appelt and Konolige, 1988] on speech acts. The emphasis there, however, is on the default assumptions that the speaker [resp. hearer] of a speech act must make about the hearer's [resp. speaker's] beliefs. There is no discussion of the ways in which agents reason about other agent's abilities to reason non-monotonically, the focus of this paper.

tions. This mutual absence of doubt is likewise best modelled by multi-agent non-monotonic reasoning.

2. Extending AEL to MANML

We begin our formalization of a Multiple-Agent Non-Monotonic Logic (which we will call MANML) by extending a standard theory of plausible reasoning to the multi-agent case. There are several candidates for such a theory: Circumscription [McCarthy, 1980], Default Logic [Reiter, 1980], Non-monotonic Logic [McDermott and Doyle, 1980], and Autoepistemic Logic (AEL) [Moore, 1985]. We choose to extend Moore's AEL to the multiple-agent case. [2] This is primarily because the semantics underlying AEL is that of belief, and when we talk about agents engaging in multiple-agent non-monotonic reasoning, we most often talk in terms of knowledge and belief. I believe that Carol believes that it takes 40 minutes to get from her home to my neighborhood; Susan believes that my bird can fly. It is therefore reasonable to formulate such a system of reasoning within a logic that makes explicit reference to an agent's beliefs.

AEL was designed to formalize how an agent reasons about his own beliefs. Sentences of AEL are defined by the following rules:
(1) if ϕ is a formula of the predicate calculus, $\phi \in$ AEL;
(2) if $\phi \in$ AEL, $L\phi \in$ AEL, where L is the standard belief operator;
(3) if ϕ and ψ are sentences of AEL, so are $\phi \wedge \psi$ and $\neg\phi$.

We say a theory T of AEL is a stable set if it obeys the following three rules:
[1] T is closed under logical consequence
[2] if $P \in T$, then $LP \in T$
[3] if $P \notin T$, then $\neg LP \in T$.

That is, AEL models a perfect reasoner who has perfect positive and negative introspection.

Since Moore only considered single agents there was no need to index the belief operator L. Since we are modelling a multiple agent world, we do. We thus state the formation rules of MANML as follows:
(1) if ϕ is a sentence of the predicate calculus, ϕ is a sentence of MANML.
(2) if ϕ is a sentence of MANML, $L_a\phi$ is a sentence of MANML, where a is a constant of the language that represents an agent
(3) if ϕ and ψ are sentences of MANML, so are $\phi \wedge \psi$ and $\neg\phi$

Once we introduce multiple agents into the theory, the stable set formation rules of AEL should no longer hold. If P is in T, we do not necessarily want to say that L_aP is in T, for any a. Just because a formula is true in a theory, we would not want to say that any agent believes that formula.

Nevertheless, we wish to get the effect of these stable set formation rules, so that agents can reason autoepistemically, and so that agents can reason about other agents reasoning autoepistemically. The simplest way to do this is to alter the stable set formation rules by adding an explicit level of indexing in the obvious way. This yields the following set of rules:

0. if $P_1, ..., P_n \in T, P_1 ... P_n \vdash Q$, then $Q \in T$.
1. if $L_aP_1, ..., L_aP_n \in T, P_1 ... P_n \vdash Q$, then $L_aQ \in T$
2. if $L_aP \in T$, then $L_aL_aP \in T$
3. if $L_aP \notin T$, then $L_a\neg L_aP \in T$

Note that rule 3. is close, but not identical, to the principle of negative introspection in theories of belief: $\neg L_aP \Rightarrow L_a\neg L_aP$.

If we also assume that agents never believe contradictions, we get the following consequences:

4. if $L_aL_aP \in T$, then $L_aP \in T$
5. if $L_a\neg L_aP \in T$, then $L_aP \notin T$

Default rules must also be indexed appropriately. Bill's belief that he would know if he had an older brother is represented as $L_{Bill}(P \Rightarrow L_{Bill}P)$, where P stands for the sentence: Bill has an older brother. Suppose $L_{Bill}P \notin T$. By 1., $L_{Bill}(\neg L_{Bill}P \Rightarrow \neg P)$. But, since $L_{Bill}P \notin T$, by 3., $L_{Bill}\neg L_{Bill}P \in T$. Thus, by 1., $L_{Bill}\neg P$. Thus, we can see that the MANML stable set formation rules allow an agent to reason from his lack of belief in a particular proposition to the fact that he does not believe a proposition.

The MANML stable set formation rules 1. - 3. were chosen so that agents in MANML could perform autoepistemic reasoning. To show that that this is the case, we must prove a formal connection between AEL's and MANML's stable set formation rules. We begin with some definitions:

Let T be a set of sentences of MANML. T is *single-indexed* if all occurrences of L are indexed by the same constant. T is *epistemically closed* if all sentences in T are of the form $L_\alpha P$ for some α and P. For example, the theory $\{L_B(P \Rightarrow L_BP)\}$ is single-indexed and epistemically closed; the theory $\{P \wedge Q \Rightarrow L_aP)\}$ is single-indexed but not epistemically closed, and the theory $\{L_ap, L_bq\}$ is not single-indexed, but is epistemically closed.

We define the following function on single-indexed theories:

Strip(T) replaces every instance of L_α in T by L.

We then have the following theorem:

[2] Moore's Autoepistemic Logic has in some sense been superseded by Konolige's extensions to it [Konolige, 1987]. We use here Konolige's first extension: his move from propositional to predicate logic. In subsequent extensions, Konolige also gave a stronger notion of groundedness that eliminates circular reasoning. We choose the first extension as a starting point because of its simplicity.

Theorem1: Let T be a single-indexed and epistemically closed collection of sentences of MANML. Let $T_\alpha = \{\phi | L_\alpha \phi \in T\}$. Then, $L_\alpha P$ is a MANML stable-set consequence of T iff P is an AE stable-set consequence of $Strip(T_\alpha)$.

The theorem follows directly from the definitions.

Note that, in restricted cases, MANML seems to permit *other* agents to reason about an individual agent's autoepistemic reasoning. Assume that the principle of negative introspection: $\neg L_a P \Rightarrow L_a \neg L_a P$ holds in T. Let Q stand for the sentence: Alex has an older brother. Now suppose that T contains the following axioms: $L_{Bill} L_{Alex}(Q \Rightarrow L_{Alex} Q))$ (Bill believes that Alex believes that if Alex had an older brother, Alex would know about it) and $L_{Bill}(\neg L_{Alex} Q)$ By negative introspection and rule 1. of MANML, $L_{Bill} L_{Alex} \neg L_{Alex} Q$, and thus, by rule 1. we get $L_{Bill} L_{Alex} \neg Q$.

It should be noticed that in the foregoing example, Bill did not really reason about Alex's autoepistemic reasoning abilities at all. He started with two beliefs about Alex: Alex's default belief that if he had an older brother, he would know about it, and that Alex didn't believe he had an older brother. Given Bill's explicit belief that Alex did not have a belief about having an older brother, he was able to conclude that Alex believed that he did not have an older brother using only the principles of negative introspection and consequential closure. But this goes against the spirit of autoepistemic reasoning. The point is to start out from the positive beliefs in one's data base, use the stable set principles to conclude that there are beliefs that one doesn't have, and to go from there to negative beliefs. One should not have to explicitly *assume* the lack of positive beliefs in order to conclude that one has negative beliefs.

Similarly, suppose Susan believes that birds typically fly. Following Konolige's [1987] suggestion for representing default rules in autoepistemic logic, adding the appropriate indexing, and doing universal instantiation, we get:

$$L_{Susan}(L_{Susan} Bird(X) \wedge \neg L_{Susan} \neg Fly(X) \Rightarrow Fly(X)),$$

where X stands for Tweety. Suppose also that Susan knows that Tweety is a bird:

$$L_{Susan} Bird(X)$$

and that James knows that Susan has these beliefs:

$$L_{James} L_{Susan}(L_{Susan} Bird(X) \wedge \neg L_{Susan} \neg Fly(X) \Rightarrow Fly(X))$$

$$L_{James} L_{Susan} Bird(X)$$

In order for James to conclude that Susan believes that Tweety flies, James must also believe that Susan does *not* believe that Tweety cannot fly. That is, there must exist in T the belief:

$$L_{James} \neg L_{Susan} \neg Fly(X)).$$

Then the desired conclusion:

$$L_{James} L_{Susan} Fly(X)$$

follows by negative introspection and consequential closure. Again, these constraints go directly against the spirit of non-monotonic reasoning. The whole point is that agents need not have explicit knowledge of the conditions that are assumed to be true by default.

The question, then, at the heart of a system of multiple-agent non-monotonic reasoning is this: How is one agent to reason about a second agent's non-monotonic reasoning abilities? What can any agent coherently conclude about the beliefs that another agent does *not* have?

It is crucial to note that the multi-agent case is not at all symmetric with the single-agent case. In the single-agent case the given theory was a complete description of the mind of some agent. In the multi-agent case, agents have at best a partial description of other agents' beliefs.

The core of our approach to modelling multi-agent non-monotonic reasoning is this: Agents reason about how other agents reason non-monotonically by making default assumptions about what these agents *do not* believe. We make two important observations: Firstly, one agent may incorrectly assume that a second agent does not believes some statement P. Thus, the first agent's default assumptions are defeasible. In this sense, MANML is very different from AEL, which as Moore pointed out, is not defeasible at all. Secondly, this strategy embodies a certain amount of arrogance. An agent who reasons about a second agent's non-monotonic reasoning process must be arrogant with respect to his beliefs about the *limitations* of the second agent's beliefs. That is, he must in some sense believe that he knows all that is important to know about the second agent's beliefs.

Even if necessary, arrogance is not an attractive quality, and in too large doses, will certainly lead to wrong conclusions. Our aim, therefore, is to limit this arrogance as much as possible. For any default rule of the form $L_a \alpha \wedge \neg L_a \beta \Rightarrow \gamma$ let us call $\neg L_a \beta$ the *negative* part of the rule, since it deals with an agent's negative beliefs. To enable multi-agent non-monotonic reasoning, we need only assume that agents are arrogant with respect to the negative parts of the default rules.

A first step towards a principle of inference for MANML might therefore be:

If an agent X believes that a second agent Y believes some default rule $L_Y \alpha \wedge \neg L_Y \beta \Rightarrow \gamma$, and X believes that Y believes α and has no reason to believe that Y believes β, X can conclude that Y believes γ.

Formally, suppose

$$L_X L_Y(L_Y \alpha \wedge \neg L_Y \beta \Rightarrow \gamma) \in T,$$

$$L_X L_Y \alpha \in T,$$

$$L_X L_Y \beta \notin T.$$

Then,

$$L_X L_Y \gamma \in T.$$

Note that the $L_Y\alpha$ part of the rule may be empty; the rule thus covers autoepistemic rules of the form: $P \Rightarrow L_X P$.
We will call the above principle the Principle of Moderate Arrogance (PMA). If Q can be inferred from a theory T using the MANML stable set principles and PMA, we say Q is a PMA-stable-set consequence of T.

It can easily be seen that the Principle of Moderate Arrogance allows us to model in a rational manner how Bill comes to conclude that Alex believes that he has no older brother. In particular, it is an instance of the following theorem:

Theorem 2: Let $T = \{L_X L_Y(P \Rightarrow L_a P)\}$ be a theory of MANML + PMA. Then $L_X L_Y \neg P$ is a PMA-stable-set consequence of T.

Often, even an arrogant agent finds it worthwhile to be more circumspect about ascribing the absence of beliefs to other agents. This is particularly the case when the arrogant agent *does* believe the negative part of some default rule. It is very often difficult to believe that someone knows less than you do - so if you have stumbled across some unexpected circumstance, you would not want to assume that other agents are ignorant of it. This is especially true if the first agent has any reason to mistrust, or fear the actions of, the second agent (e.g., in cases where agents have conflicting goals, such as enemy agents in wartime).

We call this rule of inference the Principle of Cautious Arrogance (PCA). It is formalized as follows:

Suppose
$L_X L_Y(L_Y\alpha \wedge \neg L_Y\beta \Rightarrow \gamma) \in T$,
$L_X L_Y\alpha \in T$,
$L_X L_Y\beta \notin T$, and
$L_X\beta \notin T$.
Then
$L_X L_Y\gamma \in T$.

The PCA may be too cautious at times. There are certainly cases in which one agent knows that another agent is more ignorant than he - as in the restaurant example in section 1 - and in these cases one would rather adopt PMA than PCA. In general, however, both principles are much too permissive. For example, it is reasonable to assume that virtually all agents believe that if they had an older brother, they would know about it. Nevertheless, I would not want to conclude of every new person that I meet that they do not have an older brother, simply because I do not know that they believe that they have an older brother! Yet these conclusions would be sanctioned by both the PMA and PCA.

Clearly, we need restricted versions of the PMA and PCA for different situations. There are two ways to go about formalizing these restrictions: (1) posit that arrogance is a binary relation between agents. Bill may be arrogant about Alex's knowledge but not about Susan's, if he knows Alex very well and Susan only slightly. (2) restrict the types of defaults about which agents are arrogant. In particular, it may be the case that agents do some sort of default reasoning more readily the others. The restrictions should capture these tendencies.

Much research needs to be done on both these fronts in order to develop a realistic system of multiple-agent non-monotonic reasoning. In particular, we believe that it will be instructive to look at specific domains of multiple-agent non-monotonic reasoning in which restricted versions of PMA or PCA do seem reasonable. These exercises should give us insight into the reasons the arrogant rules of inference yield intuitive results in many cases, and ought to point us toward modifying these principles into truly reasonable rules of inference.

3. Epistemic Motivated Action Theory in MANML

A promising domain for the formalization of a restricted version of PMA is that of temporal reasoning, and specifically, of temporal projection. Agents who operate in multi-agent domains must reason about other agents' abilities to predict which facts about the world stay the same, and which change. We examine below a difficult problem in temporal projection, discuss its solution in a model-preference based theory known as EMAT [Morgenstern, 1989], and show that the principle embodied in EMAT can be recast as a restricted form of PMA.

Consider, then, the following problem, which we will call the Chain Request Frame Problem: [3] Suppose Alice wants to open a safe. She knows that the combination of the safe is a sequence of three two-digit numbers, but she does not know which. That is, she knows a finite disjunction of the form: *Comb = N1 or Comb = N2 or .....* The combination of the safe is a fluent; various authorized individuals may change the combination of the safe. However, typically, the combination does not change; this is a fluent with a long persistence. Given the large number of possible combinations, it would not be wise to attempt all of them. Alice knows that Jim knows the combination of the safe, but she is not on good enough terms with him to ask him. However, she knows Susan quite well, and Susan is on good terms with Jim. Alice therefore constructs the following 4-step, multi-agent plan:

1. Alice asks Susan to ask Jim for the combination
2. Susan asks Jim for the combination

[3] The Chain Request Frame Problem is a synthesis and simplification of two variant frame problems, the Third Agent Frame Problem, and the Vicarious Planning Frame Problem, which are discussed in [Morgenstern, 1989]. These frame problems and their solutions in EMAT, were developed in a rich logical language that allowed quantification into epistemic contexts. AEL (even Konolige's extended version) and therefore MANML do not allow quantification into epistemic contexts.

3. Jim tells Susan the combination
4. Susan tells Alice the combination

Unfortunately, Alice cannot prove that this plan will work. The reason is that although Jim knows the combination at the time when Alice begins her plan, Alice does not know that he will still know the combination by the time Susan asks him for it. Frame axioms will not work: since Jim is not involved in the initial stage of the plan, he may not know what happens, and therefore will not be able to apply the frame axioms. For the same reason, neither do standard non-monotonic temporal logics ([Lifschitz, 1987], [Haugh, 1987], and [Baker and Ginsberg, 1988]). Very briefly, the reason these logics will not work is that they are based on the situation calculus, and are therefore *dense* and/or *complete* in the following sense: A theory is *dense* if there are no gaps; for any interval in the past, one always knows of some action or subaction that has occurred during that subinterval. A theory is *complete* if all actions that have occurred are known to have occurred. In cases where a particular chronicle is not dense and/or not complete for a particular agent - as in the case of Jim, above - such logics cannot be used.

Nevertheless, a system capable of commonsense reasoning should be able to conclude that Alice's plan will probably work. Most likely, the combination of the safe will not change. Jim knows this. Therefore, as long as Jim does not know of anything that would indicate that the combination has changed, he will assume that it has not changed; i.e., he will still know the combination. This is true regardless of whether Jim knows what has happened in the initial stage of Alice's plan. Alice does not know of anything that Jim knows that would indicate a change; therefore she reasons that he will know the combination when Susan asks him, and will be able to tell her.

The basic principle underlying the foregoing reasoning is that actions happen only if they have to happen, or are *motivated*. This principle has been formalized in a non-monotonic logic called Motivated Action Theory (MAT) [Morgenstern and Stein, 1988] and is given in terms of a preference criterion over models. We assume a theory of causal and persistence rules TH, and a collection of statements giving a partial description of a chronicle, called a chronicle description CD. $CD \cup TH = TI$, a particular theory instantiation. All rules are monotonic; non-monotonicity is achieved through the preference criterion. A statement is said to be *motivated* if it has to be in all models of TI (is a theorem of TI); a statement is said to be motivated with respect to a particular model if it has to be true, given rules and boundary conditions, within that particular model. More specifically, if TH contains a rule of the form

$$\alpha \wedge \beta \Rightarrow \gamma$$

where α is a conjunction of sentences of the form $True(t, Occurs(act))$, i.e., the *triggering events* of the causal rule, β contains the conditions which must be true for the triggering events to have an effect, and γ is the result of the triggering events, and it is the case that α is motivated, and β is true with respect to some particular model, then γ is *motivated* with respect to that model. We prefer models which minimize statements of the form $True(t, Occurs(act))$; i.e., in which as few unmotivated actions as possible occur. Note that this is not the same as minimizing occurrences; in particular, the two concepts are different for causal chains of events in which the triggering event is motivated.

We have demonstrated that MAT models both forward and backward temporal reasoning, and in particular have shown that it gives intuitive results for the Yale Shooting Problem and a host of related problems. [4]

EMAT extends MAT by parameterizing theory instantiations with respect to agents and times. For example, $TI(a, t1)$ describes a's beliefs at $t1$ with regard to the causal theory and the description of the chronicle that he is in; $TI(a, t1, b, t2) = TI(a, t1)(b, t2)$ describes what a at $t1$ believes b at $t2$ believes. Motivation within a parameterized theory instantiation is analogous to motivation within a standard theory instantiation; similarly, the preference criterion over models of a parameterized theory instantiation is analogous to the preference criterion over models of a standard theory instantiation. The net result is that agents always assume that other agents reason using MAT on the theory instantiations which they ascribe to them.

In the above example, EMAT allows Alice to prove that her 4-step plan will work. The theory instantiation TI(Alice,1,Jim,3) contains the statement that the combination at time 3 is identical to the combination at time 1; thus, Jim knows the combination.

EMAT provides a simple, intuitive solution to the problem of temporal projection in epistemic contexts. It is interesting to note, however, that a very basic assumption of arrogance lies at the foundation of EMAT. By using the inference rules of MAT within a parameterized theory instantiation, agents in EMAT implicitly assume that the partial characterization that they have of the other agents' theory instantiations is sufficient for their purposes. It is implicitly assumed that if $TI(b, t2)$ contained some unexpected action, then $TI(a, t1, b, t2)$ would contain this action as well. That is, agents are arrogant with respect to what they know regarding other agents' beliefs about the course of events.

In fact, it is straightforward to model the basic principle of EMAT as a restricted form of PMA. To see

[4] Specifically, MAT yields the desired results for the Bloodless Yale Shooting Problem [Morgenstern and Stein, 1988], the Stanford Murder Mystery [Baker and Ginsberg, 1988], the Waiting Can Kill You [unnamed problem in [Baker and Ginsberg, 1988], p. 5], and the Message Passing Problem [Morgenstern and Stein, 1988]

this, note that the intuition underlying MAT - actions happen only if they have to happen - can be captured by the following axiom schema of MANML:

$$L_a(\neg L_a Occurs(act) \Rightarrow \neg Occurs(act))$$

Equivalently,

$$L_a(Occurs(act) \Rightarrow L_a Occurs(act))$$

That is, a believes that if it is consistent for him to believe that an action has not occurred, then the action has not occurred. In other words, it is assumed by default that unmotivated actions do not occur.

This assumption can be made explicit in the following restricted form of the PMA, which is limited to default rules of causal reasoning. This restricted form of the PMA, (EMAT-PMA) can be stated as follows: [5]

Suppose
$L_X L_Y (L_Y \alpha \wedge \neg L_Y True(t, Occurs(act)) \Rightarrow \gamma) \in T$
$L_X L_Y \alpha \in T$
$L_X L_Y True(t, Occurs, act)) \notin T$
Then
$L_X L_Y \gamma \in T$

This gives us a powerful, but not overly permissive, inference rule for non-monotonic temporal reasoning.

Thus far, we have shown that MANML + EMAT-PMA gives identical results to EMAT for the Chaining Request Frame Problem, and the class of Yale Shooting Problems. We are currently working on a proof of the claim that MANML + EMAT-PMA is equivalent to EMAT, modulo quantification into epistemic contexts.

Naturally, EMAT-PMA models only some of the reasoning power that a genuine theory of commonsense reasoning must have. Nevertheless, the reasonableness of this inference rule suggest the possibility that a group of rules of this sort, each expressing a restriction of PMA for some sort of reasoning, is a good first step toward building a general purpose theory of multiagent non-monotonic reasoning.

4. Conclusions and Future Work

We have argued that a theory of multiple-agent non-monotonic reasoning is crucial for a realistic theory of commonsense reasoning. We have presented MANML, a logic which is capable of expressing such reasoning, and have suggested two inference rules to allow this type of non-monotonic reasoning, based on the concept of an agent's arrogance towards his knowledge of another agent's ignorance. While they are good first approximations, these rules were shown to be overly permissive. It was suggested that domain-specific restrictions of the principles of arrogance would give a more realistic theory. Finally, we demonstrated that an existing theory of temporal reasoning, which allowed for limited multiple-agent non-monotonic reasoning, could be duplicated by a restricted form of one of the principles of arrogance.

Future work includes investigating further restrictions of the PMA or PCA for specific domains of commonsense reasoning. One promising domain seems to be that of speech acts theory. Gricean theory [Grice, 1957] has argued that a mutual knowledge of convention is a prerequisite for successful communication; Russell [1987] has argued that a more realistic theory would be based on the concept of a mutual absence of doubt on the part of the speakers that they have different conventions. That is, it should be reasonable to assume that the agent with whom you are communicating shares your conventions unless you can prove otherwise. This assumption is predicated on some amount of arrogance towards the other agent; you believe that if he had different conventions (i.e., did *not* believe the accepted conventions), you would know about it. Let us introduce the operator LCL, which we define equivalent to [Cohen and Levesque, 1987] BMB (believe that it is mutually believed) operator. Then, a first pass at modelling this sort of inference rule might be:

Suppose $L_X(\neg LCL_Y(X, Y, Convention_i)) \notin T$.
Then $L_X LCL_Y(X, Y, Convention_i) \in T$.

This is clearly a restricted form of PMA. The accuracy and usefulness of this inference rule for formalizing speech acts theory is a topic for further investigation.

Finally, we plan to integrate MANML with various existing theories of multiple-agent commonsense reasoning, starting with a robust theory of planning and action. We can then test the utility of MANML on AI commonsense reasoning problems that thus far have been solvable only within a monotonic logic.

Acknowledgements

This work has benefited from the criticisms, suggestions and comments of Francisco Corrella, Ernie Davis, Hector Geffner, Benjamin Grosof, Daniel Lehmann, Lynn Stein, and Wlodek Zadrozny.

References

[Appelt and Konolige, 1988] Appelt, Douglas and Kurt Konolige: "A Practical Nonmonotonic Theory for Reasoning About Speech Acts, " *Proceedings of the 26th Conference of the ACL*, 1988

[Baker, 1989] Baker, Andrew: "A Simple Solution to the Yale Shooting Problem," *Proceedings, First Conference on Principles of Knowledge Representation and Reasoning*, 1988

[Baker and Ginsberg, 1988] Baker, Andrew and Matthew Ginsberg: "Some Problems in Temporal Reasoning," manuscript, 1988

[5] To ensure that theorems of MAT are also theorems of MANML + EMAT-PMA, we must also add the axiom of privileged access [Davis, 1990]: $L_X L_X P \Rightarrow L_X P$. This is of course a consequence of the stable set principles of MANML if we assume that agents do not believe contradictions.

[Cohen and Levesque, 1987] Cohen, Philip and Hector Levesque: "Rational Interaction as the Basis for Communication," *Proceedings, Symposium on Plans and Intentions in Communication and Discourse*, Monterey, 1987

[Davis, 1990] Davis, Ernest: *Representations of Commonsense Knowledge*, Morgan Kaufman, 1990

[Ginsberg, 1987] Ginsberg, Matthew, ed: *Readings in Nonmonotonic Reasoning*, Morgan Kaufman, Los Altos, 1987

[Grice, 1957] Grice, H.P.: "Meaning," *Philosophical Review*, 1957

[Haugh, 1987] Haugh, Brian: "Simple Causal Minimizations for Temporal Persistence and Projection," *Proceedings, AAAI 1987*

[Konolige, 1987] Konolige, Kurt: "On the Relation Between Default Theories and Autoepistemic Logic," *Proceedings, IJCAI 1987*

[Lifschitz, 1987] Lifschitz, Vladimir: "Formal Theories of Action," *Proceedings, IJCAI 1987*

[McCarthy, 1980] McCarthy, John: "Circumscription," *Artificial Intelligence*, Vol. 13, 1980

[McDermott and Doyle, 1980] McDermott, Drew and Jon Doyle: "Non-monotonic Logic I," *Artificial Intelligence*, Vol. 13, 1980

[Moore, 1985] Moore, Robert: "Semantical Considerations on Nonmonotonic Logic," *Artificial Intelligence*, Vol.25, 1985

[Morgenstern, 1989] Morgenstern, Leora: "Knowledge and the Frame Problem," Workshop on the Frame Problem, Pensacola, 1989. To appear in: Kenneth Ford and Patrick Hayes, eds: *Advances in Human and Machine Cognition*, Vol. I: The Frame Problem in Artificial Intelligence, JAI Press, Greenwich, 1990

[Morgenstern and Stein, 1988] Morgenstern, Leora and Lynn Andrea Stein: "Why Things go Wrong: A Formal Theory of Causal Reasoning," *Proceedings, AAAI 1988*

[Perrault, 1987] Perrault, Ray: "An Application of Default Logic to Speech Act Theory," *Proceedings, Symposium on Intentions and Plans in Communication and Discourse,* Monterey, 1987

[Reiter, 1980] Reiter, Ray: "A Logic for Default Reasoning," *Artificial Intelligence*, Vol. 13, 1980

[Russell, 1987] Russell, Stuart: "Rationality as an Explanation of Language?," *Behavioral and Brain Sciences*, Vol. 10, 1987