Introspection and Projection in Reasoning about Other Agents

Jacques Wainer
Departamento de Ciência da Computação
UNICAMP

DCC—IMECC
Universidade Estadual de Campinas
Caixa Postal 6065
13081-970 Campinas - SP
Brasil

wainer@dcc.unicamp.br

Abstract

This paper develops the formal aspects of a new approach to reasoning about the knowledge of other agents. It is based on the principles of introspection, by which an agent is aware of the inferences he makes, and projection, by which an agent assumes that other agents have the same inference abilities as himself. The paper develops a logic that incorporates these principles and proves the soundness of such logic.

1 Introduction

Reasoning about the knowledge of other agents is of fundamental importance if a system is to interact in social situations. An important part of communication is based on inferring and attributing knowledge to the other agents in a conversation. For example, if both agents A and B are at agent A's house, and after hearing a dog's bark, A can say:

> *That* dog is driving me crazy at night.

If A expects B to understand his utterance, he must believe that B knows that dogs (usually / sometimes / can) bark, and that if a dog can bark it can bark at night, and that barking may keep one awake, and that staying awake may drive one crazy. A must also believe that B can figure out that the dof refered by the pronoum that in the utterance is the one that just barked (because there are no other dogs in the context). Finally, A must believe that B can put all that together, and conclude that the dog that just barked is keeping A awake at night and thus driving him crazy, by virtue of its barking.

Of course, the reasoning that lead to the conclusion that the dog is keeping A awake by virtue of its barking is a very complex form of plausible reasoning, which will not be addressed by this paper. The example was meant to show that attributing knowledge to other agents and reasoning what these agents can conclude from this knowledge are common aspects of the communication among agents. In this paper we will be dealing with logical monotonic reasoning and reasoning about others ability to performer this kind of reasoning. But examples of reasoning about other agents ability to perform logical reasoning are so obvious that they do not make it salient that one is in fact reasoning about other agent's knowledge and reasoning power. For example, if A knows that B knows that fish are not mammals, and A tells B that he has a golden fish in an aquarium, then A can conclude that B knows (or should be able to know) that the golden fish is not a mammal. This is an example of the form of reasoning that we will be dealing with in this paper.

1.1 The main intuition

How is A able to reason about B's knowledge? Our main intuition is that A performs (or could perform) this reasoning about the knowledge of another agent through a mechanism of introspection and projection. Agent A could perform the following reasoning: "what would I conclude given the same knowledge I think B has" (this is the introspection step), and then attribute these conclusions to B (this is the projection step). In fact our main intuition is that humans do use introspection and projection when reasoning about other agents knowledge, and that an artificial agent could be implemented using these same principles. These ideas were originally proposed in [MWC91].

The goal of this paper is to show in a formal way that the principle of introspection and projection could indeed serve as foundation for an artificial agent that reasons about other agents' knowledge. We will develop a logic based on these principles and compare it to other logics that have been developed to model the knowledge of many agents.

1.2 Internal and external logics for belief

McCarthur [McA88] discusses the distinction between an internal and an external logic for belief. An **internal logic** assumes a particular reasoner's the point of view: a formula is true if the agent "knows" it, and the inferences in the logic are a model or an abstraction of the agent's own reasoning process. Therefore, asserting

$$bird(tweety)$$

in an internal logic states that the agent who is being modeled by the logic knows that Tweety is a bird. Internal logics has been used within AI to describe nonmonotonic forms of reasoning (for example [Rei80, McC80, Del88, Moo85], representing the main branches of the formalization to nonmonotonic reasoning), with the exception of the work by [Lev90] which describes nonmonotonicity from an external point of view. Thus the usual nonmonotonic inference rule:

$$\frac{\text{Usually bird fly} \\ \text{Tweety is a bird}}{\text{Tweety flies}}$$

is to be understood from an internal point of view: if you know that birds usually fly, and you know that Tweety is a bird, and that is all you know, then conclude that Tweety flies.

On the other hand, an external logic assumes a “reality” point of view, and asserting a formula to be true means it holds in reality. In this case, to assert that the agent i knows something, one needs an operator that explicitly refers to the knowledge of that agent ($\mathbf{B}_i$). To assert that agent i knows that Tweety is a bird, one would assert the formula:

$$\mathbf{B}_i bird(tweety)$$

In this paper there will be a close interplay between internal logics and external logics for belief. The reason is that the mechanisms of introspection and projection are really models of the reasoning process of an agent, and therefore should be captured within an internal logic framework. On the other hand, formulas expressed in an external logic are easier to understand, and external logics are more expressive that internal logics, since they can represent the concept of lack of knowledge ($\neg\mathbf{B}$), which cannot be expressed in an internal logic. Furthermore, we would like to compare the logic developed here with some of the more standard logics that also model the knowledge of many agents. The only of such logics we are aware of are the ones based on multi-operators modal logics, for example the ones discussed in /citeguide, and they are all external logics. Thus, in this paper we will alternate between describing the intuitions behind the introspection and projection principle in an internal logic, and formalizing them in a external logic.

1.3 The language

We will use a modal propositional language to represent knowledge or better, belief. $\mathbf{B}_i q$ represents the statement that the agent i believes that q is true. $\mathbf{P}_i q$ represents the statement that for all i knows, q is possible. As usual, $\mathbf{P}_i$ is taken as the dual of $\mathbf{B}_i$, thus the following holds:

$$\mathbf{P}_i q \leftrightarrow \neg\mathbf{B}_i \neg q$$

When we have to switch to an internal logic to describe an agent's reasoning from his own point of view, we will use the symbol “$\mid\!\sim_i$” to represent the consequence relation in the internal logic of agent i. Thus,

$$\alpha \mid\!\sim_i \beta$$

states that if agent i knows α then he would also know β.

1.4 Modal logics of knowledge for multiple agents

The usual logic for knowledge of multiple agents is defined by extending a modal logic of knowledge for a single agent. Possible base logic for such an extension are S4, S5 and weak-S5 (or technically KD45). In this paper we will use the modal logic weak-S5 as the underlying modal logic of knowledge.

We call the extension of weak-S5 to multiple agents as the logic n-KD45 (after n agents, weak-S5 underlying modal logic). n-KD45 can be captured by the following axioms and deduction rules, where $\vdash_n$ is the consequence relation in the logic n-KD45.

Axioms:

$$\begin{array}{rll}
\text{(Prop)} & \vdash_n & \alpha \quad \text{if} \quad \alpha \ \text{is a tautology of propositional logic} \\
\text{(K)} & \vdash_n & \mathbf{B}_i\alpha \wedge \mathbf{B}_i(\alpha \rightarrow \beta) \rightarrow \mathbf{B}_i\beta \quad \text{for all agents } i \\
\text{(D)} & \vdash_n & \mathbf{B}_i\alpha \rightarrow \neg\mathbf{B}_i\neg\alpha \\
\text{(4')} & \vdash_n & \mathbf{B}_i\alpha \leftrightarrow \mathbf{B}_i\mathbf{B}_i\alpha \\
\text{(5')} & \vdash_n & \neg\mathbf{B}_i\alpha \leftrightarrow \mathbf{B}_i\neg\mathbf{B}_i\alpha \\
\text{(And)} & \mid\!\sim & \mathbf{B}_i\alpha \wedge \mathbf{B}_i \leftrightarrow \mathbf{B}_i(\alpha \wedge \beta)
\end{array}$$

Inference rules

$$\text{(MP)} \qquad \frac{\alpha \qquad \alpha \rightarrow \beta}{\beta}$$

$$\text{(NEC)} \qquad \frac{\vdash_n \alpha}{\mathbf{B}_i\alpha} \quad \text{for all agents } i$$

We express the NEC inference rule somewhat different then the usual. As it stands above, it states that if α *is an theorem*, then one can conclude $\mathbf{B}_i\alpha$, for all agents i.

The axioms above are not minimal for the logic n-KD45. In particular the axioms $4'$ and $5'$ could be written as just an implication and the axiom **And** is derivable from the others. But this formulation will be closer to the one we will develop for the logic based on introspection and projection.

2 A logic of introspection and projection

We will now propose what we believe is a more plausible mechanism to reason about the knowledge of other agents. The mechanisms are introspection and projection. By introspection we mean both the awareness of the agent's own knowledge and the awareness of his own reasoning process. By projection we mean that the agent's belief that all other agents have the same reasoning abilities as himself, although they may have different knowledge. Thus, if the agent knows that he would be able to derive a piece of knowledge from some set of assumptions, then by using projection, he can assume that other agents faced with the same set of assumption would derive the same conclusion.

If agent 0 is able to deduct β from α, then the introspection principle allow us to claim that agent 0 would be able to know that β is deductible from α. The projection principle would then state that one agent 0 is aware that β is deducible from α he would also believe that if any other agent, say agent 1, knows α then agent 1 should also know β.

We will call the logic that embodies the introspection and projection principles as the logic l-KD45, after "local, n agents, weak-S5." The derivability relation in that logic will be denoted by $\vdash_l$.

2.1 Introspection

We will assume that the agents have both positive and negative introspection of their knowledge. Formally this is captured by the usual axioms of standard modal logic:

$$\begin{array}{lll} (4) & \vdash_l & \mathbf{B}_i\alpha \rightarrow \mathbf{B}_i\mathbf{B}_i\alpha \\ (5) & \vdash_l & \neg\mathbf{B}_i\alpha \rightarrow \mathbf{B}_i\neg\mathbf{B}_i\alpha \end{array}$$

The symbols "4" and "5" are the usual name of the axioms above in the modal literature.

The second form of introspection is more interesting. It reflects the fact that the agent is aware of his own reasoning process. If there is a derivation from α to β in the agents internal reasoning process, then the agent is aware of such derivation. In a non-formal way, if there is a derivation $\alpha \mid\!\sim_i \beta$, then the agent is aware of it: $\mathbf{B}_i(\alpha \mid\!\sim_i \beta)$. Of course this last formula is not syntactically well defined, and it mixes both the internal and external interpretation of the formulas, but it captures the intuition behind reasoning introspection.

The main assumption behind reasoning introspection is that the agent is able to collect (or remember) all assertions used in a deduction. If the agent succeed in deriving β, he can recall all propositions from his knowledge base that were used in that deduction. For example, if in proving β, the agent had to use only the propositions $\alpha_1, \alpha_2, \alpha_3$ asserted in his knowledge base, then the agent will remember this. Using the incorrect notation mentioned above, we would have $\mathbf{B}_i(\alpha_1, \alpha_2, \alpha_3 \mid\!\sim \beta)$.

The main aspect in this assumption is not that the agent is able to construct a "minimal set of premises," that is, that the premise p in $p \mid\!\sim q$ is in some way minimal, but that the agent is able to recall all premises used in the derivation, that there will not be any forgotten premises.

The informal relation "if $p \mid\!\sim_i q$ then $\mathbf{B}_i(p \mid\!\sim_i q)$ is captured within the formalism of l-KD45 as an inference rule:

$$\frac{\vdash_l \mathbf{B}_i\alpha \rightarrow \mathbf{B}_i\beta}{\mathbf{B}_i(\mathbf{B}_i\alpha \rightarrow \mathbf{B}_i\beta)}$$

The inference rule states that if $\mathbf{B}_i\alpha \rightarrow \mathbf{B}_i\beta$ is a theorem then i will know it. In other words, given no other non-logical assumption, if i knowing α implies i knowing β, then i will know that knowing α implies knowing β.

For technical reasons we will like possibility of having a conjunction as the antecedent of implication above. Therefore, the inference rule that really corresponds to the reasoning principle is

$$\text{(RI)} \qquad \frac{\vdash_l \mathbf{B}_i\alpha \wedge \mathbf{B}_i\gamma \rightarrow \mathbf{B}_i\beta}{\mathbf{B}_i(\mathbf{B}_i\alpha \wedge \mathbf{B}_i\gamma \rightarrow \mathbf{B}_i\beta)}$$

2.2 Projection

The projection principle states that if the agent knows that he is able to perform a certain deduction, then he knows that other agents are also able to do it. Again in an informal way:

$$\text{if } \alpha \mid\!\sim_i \beta \text{ then } \mathbf{B}_i(\alpha \vdash_j \beta) \qquad (1)$$

But the informal statement in (1) does not capture all of the issues about projection. One problem, for example is that either α or β may contain references to i's knowledge. For example, if

α is a proposition p, and β is $\mathbf{B}_i\neg\mathbf{B}_i\neg p$, then applying the projection principle as stated in (1), one would conclude:

$$\text{since} \quad p \mathrel{|\!\sim_i} \mathbf{B}_i\neg\mathbf{B}_i\neg p \quad \text{then} \quad \mathbf{B}_i(p \vdash_j \mathbf{B}_i\neg\mathbf{B}_i\neg p) \tag{2}$$

But (2) does not capture the idea of the projection principle: from the fact that i can deduce something about his knowledge does not mean that j should be able to deduce the same about i's knowledge. The spirit of the projection principle is that i believes that j can perform the same deduction, but in this case about j's own knowledge. Thus, the desired inference is not the one in (2), but:

$$\text{since} \quad \mathbf{B}_i(p \mathrel{|\!\sim} \mathbf{B}_i\neg\mathbf{B}_i\neg p) \quad \text{then} \quad \mathbf{B}_i(p \vdash_j \mathbf{B}_j\neg\mathbf{B}_j\neg p)$$

The conclusion in the statement $p \mathrel{|\!\sim} \mathbf{B}_i\neg\mathbf{B}_i\neg p$ should not be interpreted as a proposition about i's knowledge, but a proposition about the reasoner's own knowledge, which first is about i himself, but after the projection becomes about the projected agent, in this case j. In other words, the projection principle states that i believes that j is able to perform the same sequence deductions i did, but relative to j own knowledge.

To achieve this relativization of knowledge we need a syntactic operation that changes the indices in the $\mathbf{B}$ operator. This syntactic operation is called **substitution** and is denoted as $|^i_j$, which means substitute $\mathbf{B}_i$ by $\mathbf{B}_j$.

The substitution operation is defined recursively:

$$\begin{array}{rcl}
p|^i_j & = & p \quad \text{if } p \text{ is a propositional symbol} \\
(\alpha \wedge \beta)|^i_j & = & \alpha|^i_j \wedge \beta|^i_j \\
(\neg\alpha)|^i_j & = & \neg(\alpha|^i_j) \\
(\alpha \rightarrow \beta)|^i_j & = & (\alpha|^i_j) \rightarrow (\beta|^i_j) \\
(\mathbf{B}_i\alpha)|^i_j & = & \mathbf{B}_j(\alpha|^i_j) \\
(\mathbf{B}_x\alpha)|^i_j & = & \mathbf{B}_x\alpha \quad \text{for } x \neq i
\end{array}$$

The last line of the recursive definition disables the substitution of $\mathbf{B}_i$ by $\mathbf{B}_j$ if that operator is inside the scope of another operator. This is to capture the intuition that if $\mathbf{B}_i$ appears within the scope of say $\mathbf{B}_r$ then it no longer refer to the "knowledge of the reasoner" (which happens to be i but gets substituted by j by projection), but it indeed refers to the "knowledge of the agent" i. Thus, for example if the formula $\mathbf{B}_i\mathbf{B}_r\neg\mathbf{B}_i p$ is the conclusion of some reasoning performed by i, then projecting that reasoning to j would allow i to conclude that $\mathbf{B}_j\mathbf{B}_r\neg\mathbf{B}_i p$ if j knew the premises that support the first conclusion. In other words, the conclusion is about r's beliefs in i's lack of knowledge of p, and projection would allow j to conclude the same thing.

The principle of projection is captured by the following inference rule, which also include for technical reasons a conjunction as the antecedent.

$$(\text{P}) \quad \frac{\vdash \mathbf{B}_i\alpha \wedge \mathbf{B}_i\gamma \rightarrow \mathbf{B}_i\beta}{\mathbf{B}_i(\mathbf{B}_j\alpha|^i_j \wedge \mathbf{B}_j\gamma|^i_j \rightarrow \mathbf{B}_j\beta|^i_j)}$$

The inference rule P captures the intuition that agent i believes that whatever he can conclude given his knowledge, others will be able to conclude given the same knowledge.

There is another possible formalization for the projection principle. As it stands above, P in some way also incorporates RI. One could separate the two so that projection would only be applied after the agent had performed RI on his own reasoning. The new projection which we call P' is based on the intuition:

$$\text{if } \mathbf{B}_i(\alpha \mathrel{|\!\sim}_i \beta) \text{ then } \mathbf{B}_i(\alpha|^i_j \vdash_j \beta|^i_j)$$

or in a formal way

$$(\text{P'}) \quad \frac{\vdash \mathbf{B}_i(\mathbf{B}_i\alpha \wedge \mathbf{B}_i\gamma \rightarrow \mathbf{B}_i\beta)}{\mathbf{B}_i(\mathbf{B}_j\alpha|^i_j \wedge \mathbf{B}_j\gamma|^i_j \rightarrow \mathbf{B}_j\beta|^i_j)}$$

The logic that incorporates the P' inference rule instead of P will be called l_2-KD45. In this paper we will only deal with the logic l-KD45, since the proof of the equivalent of theorem 1 for the logic l_2-KD45 still elude us.

2.3 Other axioms about the agents

There are a few other axioms that capture the agent's reasoning process. First we must assume that the agent's reasoning is close under implication. This is captured by the usual axiom:

$$(\text{K}) \quad \vdash_l \quad \mathbf{B}_i\alpha \wedge \mathbf{B}_i(\alpha \rightarrow \beta) \rightarrow \mathbf{B}_i\beta$$

The axiom of l-KD45 above corresponds to the following principle expressed in i's internal logic:

$$\alpha \wedge (\alpha \rightarrow \beta) \mathrel{|\!\sim}_i \beta$$

Another characteristic of the agent is that his beliefs are consistent. This is captured by the axiom:

$$(\text{D}) \quad \vdash_l \quad \mathbf{B}_i\alpha \rightarrow \neg\mathbf{B}_i\neg\alpha$$

which corresponds to the following principle in i's internal logic:

$$\alpha \not\mathrel{|\!\sim}_i \neg\alpha$$

The next pair of axioms is derived from the privileged access principle. This principle states that the agent is always correct about its own knowledge, and is the reverse of the knowledge introspection principles. They are

$$(\text{4r}) \quad \vdash_l \quad \mathbf{B}_i\mathbf{B}_i\alpha \rightarrow \mathbf{B}_i\alpha$$
$$(\text{5r}) \quad \vdash_l \quad \mathbf{B}_i\neg\mathbf{B}_i\alpha \rightarrow \neg\mathbf{B}_i\alpha$$

where 4r and 5r stand for the reverse of the axioms 4 and 5.

Also the agent is able conjoint two pieces of information into a single clause and transform a conjoined clause into two separate believed clauses.

$$(\text{And}) \quad \vdash_l \quad \mathbf{B}_i\alpha \wedge \mathbf{B}_i\beta \leftrightarrow \mathbf{B}_i(\alpha \wedge \beta)$$

Finally we will assume that the agents are competent (and complete) propositional reasoners, that it they know all propositional tautologies. This is captured as the axiom

$$(\text{K-prop}) \quad \vdash_l \quad \mathbf{B}_i\alpha \quad \text{if } \alpha \text{ is a propositional tautology, for all } i$$

2.4 Other axioms of the logic

Because the logic l-KD45 is an external logic, it also has to reflect the rules of reality, independent of the agents. This means that all propositional tautologies must be axioms of the logic, and modus ponens must a inference rule. Formally:

$$\text{(Prop)} \quad \vdash_l \quad \alpha \quad \text{if } \alpha \text{ is a propositional tautology}$$

and

$$\text{(MP)} \qquad \frac{\alpha \qquad \alpha \rightarrow \beta}{\beta}$$

2.5 Summary

The axioms for the logic l-KD45 are repeated below, where the axioms 4 and 4r have been joint together and 5 and 5r has also been joint together.

$$\begin{array}{rll}
\text{(Prop)} & \vdash_l & \alpha \quad \text{if } \alpha \text{ is a propositional tautology} \\
\text{(K-prop)} & \vdash_l & \mathbf{B}_i\alpha \text{ if } \alpha \text{ is a propositional tautology, for all } i \\
\text{(K)} & \vdash_l & \mathbf{B}_i\alpha \wedge \mathbf{B}_i(\alpha \rightarrow \beta) \rightarrow \mathbf{B}_i\beta \\
\text{(D)} & \vdash_l & \mathbf{B}_i\alpha \rightarrow \neg\mathbf{B}_i\neg\alpha \\
\text{(4')} & \vdash_l & \mathbf{B}_i\alpha \leftrightarrow \mathbf{B}_i\mathbf{B}_i\alpha \\
\text{(5')} & \vdash_l & \neg\mathbf{B}_i\alpha \leftrightarrow \mathbf{B}_i\neg\mathbf{B}_i\alpha \\
\text{(And)} & \vdash_l & \mathbf{B}_i\alpha \wedge \mathbf{B}_i\beta \leftrightarrow \mathbf{B}_i(\alpha \wedge \beta)
\end{array}$$

and the inference rules:

$$\begin{array}{rc}
\text{(MP)} & \dfrac{\alpha \quad \alpha \rightarrow \beta}{\beta} \\
\text{(RI)} & \dfrac{\vdash_l \mathbf{B}_i\alpha \wedge \mathbf{B}_i\gamma \rightarrow \mathbf{B}_i\beta}{\mathbf{B}_i(\mathbf{B}_i\alpha \wedge \mathbf{B}_i\gamma \rightarrow \mathbf{B}_i\beta)} \\
\text{(P)} & \dfrac{\vdash_l \mathbf{B}_i\alpha \wedge \mathbf{B}_i\gamma \rightarrow \mathbf{B}_i\beta}{\mathbf{B}_i(\mathbf{B}_j\alpha|^i_j \wedge \mathbf{B}_j\gamma|^i_j \rightarrow \mathbf{B}_j\beta|^i_j)}
\end{array}$$

The logic l-KD45 models the knowledge of a group of agents where each one is a complete propositional reasoner, have both positive and negative introspection, and privileged access to their own knowledge, but more important they all reason about others using reasoning introspection and projection.

It is also very likely that the set of axioms above are not minimal, and that a some of the axioms above are in fact derivable form the others. This is not an important concern in this paper.

3 Soundness l-KD45

The logic l-KD45 is a strictly syntactic logic: there is no semantic model. Therefore, its soundness will be ascertained in relation with n-KD45. That is all conclusions derived in l-KD45 can also be derived in n-KD45. This is stated as the theorem below.

Theorem 1 *if* $\vdash_l\alpha$ *then* $\vdash_n \alpha$

Proof 1. By induction on the size of the proof in l-KD45. A proof in l-KD45is a sequence of formulas $\beta_1, \beta_2, \ldots . \beta_{n-1}, \alpha$, where each formula in the sequence is either an instance of the axioms of l-KD45 (Prop, K-prop, K,D,4',5', And), or the result of the application of one of the inference rules (MP, RI, P).

- **Base case**: proof of size 1 ($n = 1$). Thus α must be an instance of an axiom of l-KD45. But the axioms **Prop, K, D, 4', 5', And** of l-KD45 are also axioms of n-KD45, and thus, α is also a theorem of n-KD45. If α is an instance of **K-prop**, then it can be shown to be a theorem of n-KD45 by an instance of n-KD45's **Prop** and an application of NEC.

- **Inductive case**: the claim is true for all formulas that can be derived by proof sequences of length equal or shorter than $n - 1$. In this case, α can be a) an instance of an axiom, b) the result of an application of MP, c) the result of applying RI, or d) the result of applying P. Let discuss all these cases.

- **a)** If α is an instance of an axiom of l-KD45, then, by the arguments in the base case of this proof, α is also a theorem of n-KD45.

- **b)** If α is the result of applying MP, then there are two formulas in the proof sequence β_r and β_s such that $\beta_s = \beta_r \rightarrow \alpha$. By the inductive assumption both β_r and $\beta_r \rightarrow \alpha$ are theorems of n-KD45, and thus by applying MP (in the logic n-KD45) one concludes that α is also a theorem of n-KD45.

- **c)** If α is the result of applying RI, then some formula β_r in the proof sequence is of the form $\mathbf{B}_i\gamma \rightarrow \mathbf{B}_i\beta$ and this formula is a theorem of l-KD45.[1] Moreover, α itself is of the form $\mathbf{B}_i(\mathbf{B}_i\gamma \rightarrow \mathbf{B}\beta)$. By the inductive assumption, $\mathbf{B}_i\gamma \rightarrow \mathbf{B}\beta$ is itself a theorem of n-KD45. Applying NEC on this formula one concludes that $\mathbf{B}_i(\mathbf{B}_i\gamma \rightarrow \mathbf{B}\beta)$ is a theorem of n-KD45.

- **d)** If α is the result of applying P, then some formula β_r in the proof sequence is of the form $\mathbf{B}_i\gamma \wedge \mathbf{B}_i\gamma_2 \rightarrow \mathbf{B}_i\beta$, and α itself is of the form $\mathbf{B}_i(\mathbf{B}_j\gamma|_j^i \wedge \mathbf{B}_j\gamma_2|_j^i \rightarrow \mathbf{B}_j\beta|_j^i)$. We will need to use a lemma that is proven in Appendix A. The lemma states that if β is a theorem of n-KD45, so is $\beta|_j^i$. By the induction assumption, $\mathbf{B}_i\gamma \wedge \mathbf{B}_i\gamma_2 \rightarrow \mathbf{B}_i\beta$ is a theorem of n-KD45. By the lemma, $(\mathbf{B}_i\gamma \wedge \mathbf{B}_i\gamma_2 \rightarrow \mathbf{B}_i\beta)|_j^i$ is also a theorem of n-KD45. Applying NEC to this last theorem, one concludes that $\mathbf{B}_i(\mathbf{B}_i\gamma \wedge \mathbf{B}_i\gamma_2 \rightarrow \mathbf{B}_i\beta)|_j^i$ is also a theorem of n-KD45.

  ∎

3.1 Completeness of l-KD45

We cannot at the moment evaluate the completeness of l-KD45 in comparison with n-KD45. That is, we would like to find out which subset of the theorems of n-KD45 can be proven in l-KD45, and if this subset can be used to model interesting phenomena. In particular, we know that l-KD45 does not prove the whole set of theorems of n-KD45because it cannot project negative knowledge. For example

$$\mathbf{B}_i(\neg\mathbf{B}_j\alpha \rightarrow \mathbf{B}_j\neg\mathbf{B}_j\alpha)$$

is a theorem of n-KD45(applying NEC to an instance of 5'), but it seems that the formula cannot be derived in l-KD45, because there is no way of projecting negative knowledge ($\neg\mathbf{B}_i$). It is not

[1] This fact is not important for the proof here since we are only considering proofs without premises.

clear yet whether one would like to extend the projection and introspection principles to include negative knowledge, since it cannot be expressed in an internal logic.

3.2 Soundness of other logics

Other logics can be created based on other underlying modal logics, for example S4 and S5, but still using the principles of projection and introspection. l-S5 would have an extra axiom

$$\text{(T)} \quad \vdash_{l-S5} \quad \mathbf{B}_i\alpha \rightarrow \alpha$$

and l-S4 would also have the axiom **T** above and would not have the axiom **5'**.

Each of these two logics would also be sound with respect to their corresponding multi-agent modal logic. That is

$$\begin{array}{lll} \text{if } \vdash_{l-S4} \alpha & \text{then} & \vdash_{n-S4} \alpha \\ \text{if } \vdash_{l-S5} \alpha & \text{then} & \vdash_{n-S5} \alpha \end{array}$$

The proof of such claims follows closely the proof of theorem 1 above.

4 Conclusions

In this paper we present a new approach to reasoning about the knowledge of other agents, based on the principles of introspection and projection. We then develop a logic based on these principles and prove that the logic is consistent.

It is not easy to compare the logic l-KD45 with other approaches in terms of its "usability," or in other words, how easy is it to prove statements about the knowledge of other agents in the logic. The logic l-KD45 is intended as theoretical stepping stone of stronger logics that based on the idea that to reason about the knowledge of somebody else, the agent would "put himself on the other's shoes" and use his own reasoning process to derive conclusions that would be then attributed to the other agent. In this paper we made the hypothesis that this "putting oneself in the other's shoe" is (perhaps partially) captured by the introspection and the projection principles.

The research reported here is only starting at this point, and need some future work. An area that need more work is is the study of the completeness of the logic in relation to n-KD45, and whether the logic is able to model interesting phenomena. A second area of research is an internal characterization of the logic l-KD45. The external characterization developed here is useful in proving the properties of the logic, but cannot be used as the specification for constructing an artificial reasoner, which should be the final goal of the research.

Finally, the most exciting line of research is extending the formalism to deal with nonmonotonic reasoners. Such extension would be a interesting tool for modeling the knowledge of nonmonotonic agents [Wai92]. The principle of introspection does not depend on the fact that the agent is monotonic and thus, in principle, it could apply even if the derivation "$\mid\!\sim_i$" in $\alpha \mid\!\sim_i \beta$ is a nonmonotonic inference.

5 Appendix 1: A lemma on n-KD45

Lemma 2 *If α is a theorem of n-KD45, so is $\alpha|^i_j$.*

Proof 2. By induction on the size of the proof of α. The proof that α is a theorem is a sequence of formulas $\beta_1, \beta_2, \ldots, \beta_r, \alpha$, where each formula is an instance of an axiom, an application of modus ponens on two formulas appearing before in the sequence, or an application of necessitation on a formula that appears before in the sequence.

- **Base case:** the proof is of size 1, that is, α is an instance of an axiom. In this case, α could be an instance of the axioms prop-taut, K, D, 4, or 5, in page 4. If α is an instance of prop-taut, so is $\alpha|^i_j$ since the substitution operation does not change the propositional connectives of the formula. If α is a instance of K, that is, of the form $\mathbf{B}_x A \wedge \mathbf{B}_x(A \rightarrow B) \rightarrow \mathbf{B}_x B$, then $\alpha|^i_j$ will be the same formula if $x \neq i$ and it will be $\mathbf{B}_j A|^i_j \wedge \mathbf{B}_j(A|^i_j \rightarrow B|^i_j) \rightarrow \mathbf{B}_j B|^i_j$, which is also an instance of K. Similarly for instances of D, 4, and 5.

- **Inductive case:** The claim is true for all theorems that can be proven by proof sequences of length $n-1$ or shorter. The proof of α is the sequence $\beta_1, \beta_2, \ldots, \beta_{n-1}, \alpha$. In this proof α can be a) an instance of an axiom, b) an application of MP on two preceding formulas, or c) an application of NEC on a preceding formula.

- **a)** if α is an instance of an axiom then by the arguments in the base case of this proof, so is $\alpha|^i_j$.

- **b)** if α is an application of MP, then there are two formulas β_r and β_s in the sequence of the form β_r and $\beta_s = \beta_r \rightarrow \alpha$. By the inductive claim, both $\beta_r|^i_j$ and $\beta_r|^i_j \rightarrow \alpha|^i_j$ are theorems of n-KD45. Then applying MP on these theorems, one concludes that $\alpha|^i_j$ is also a theorem.

- **c)** if α is an application of NEC, then α is of the form $\mathbf{B}_x\beta_r$, and there is β_r in the proof sequence. If $x \neq i$ then $\alpha|^i_j$ is the same as α. If $x = i$, then $\alpha|^i_j$ is $\mathbf{B}_j\beta_r|^i_j$. By the inductive argument, $\beta_r|^i_j$ is a theorem of n-KD45, and thus by applying NEC, one concludes that $\mathbf{B}_j\beta_r|^i_j$ is a theorem of n-KD45.

  ∎

References

[Del88] J. P. Delgrande. An approach to default reasoning based on a first-order conditional logic: Revised report. *Artificial Intelligence*, 36:63–90, 1988.

[Lev90] Hector J. Levesque. All I know: A study in autoepistemic logic. *Artificial Intelligence*, 42:263–310, 1990.

[McA88] Gregory L. McArthur. Reasoning about knowledge and belief: a survey. *Computational Intelligence*, 4:223–243, 1988.

[McC80] John McCarthy. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980.

[Moo85] R. C. Moore. Semantical considerations on non-monotonic logics. *Artificial Intelligence*, 28:75–94, 1985.

[MWC91] A. Maida, J. Wainer, and S. Cho. A syntactic approach to introspection and reasoning about the beliefs of other agents. *Fundamenta Informaticae Journal*, 15(3 and 4):333–356, 1991. Special Issue on Logics for Artificial Intelligence.

[Rei80] Ray Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[Wai92] Jacques Wainer. Extending circumscription into modal domains. In *Proceedings of the 10th Meeting of the AAAI*, pages 648–653. AAAI/MIT Press, 1992.