# Reasoning about another agent through empathy

Jacques Wainer

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

wainer@dcc.unicamp.br

**Abstract**

This paper formalizes the idea of reasoning about the knowledge of another agent by "putting oneself on the other's shoes," and then attributing to the other agent the conclusions that follows from that. We call this form of reasoning empathy. We formalize empathy for monotonic knowledge and show that the same principles would apply to nonmonotonic knowledge, that is reasoning about the knowledge of nonmonotonic agents.

## 1 Introduction

How can one agent reason about another agent's beliefs? One approach is to describe a logic of belief for many agents (for example the modal logic of knowledge KD45 [HM92]) and implement a theorem prover in that logic.

Another approach, which seems closer to how humans actually do it, is to put oneself in the other agent's shoe and figure out what one would believe if one knew exactly what we believe the other agent knows. For example, if I think Ralph believes that all mammals live on land then I think that Ralph would believe that whales are not mammals, because that would be the conclusion I would derive if I believed that all mammals live on land, and that whales live in the water.

Of course this reasoning can go even further. If I think that Ralph thinks that I believe that all mammals live on land, that I can conclude that Ralph thinks that I believe that whales are not mammals, because, as the previous example showed, if I believed that about Ralph, I would conclude that he believes that whales are not mammal.

We call this mode of reasoning **empathy**[1]. Empathy can be used to reason about another agent's goals and plans (what would I do if I were Ralph and I wanted to go from his home in New York to a hotel in Paris?), feelings (what would I feel if I were Ralph, and I had I cat I loved, and it died?), and perception and identification (what would I think if I were Ralph and I didn't know that Sue has a twin sister and I met the twin sister in a party?). We will limit the scope of this paper to reasoning about beliefs.

## 1.1 This paper

The goal of this paper is to formalize empathy into a logic and to show that the logic is consistent. In this paper we will concentrate on the formal aspects of monotonic reasoning about knowledge and show that this empathic reasoning, when formalized yields a consistent logic, that in some ways can be compared to more standard logics of knowledge. Then, we will speculate on how to extend the empathy approach to reasoning about more than two agents, to reason about other modalities besides knowledge, and to reason about the knowledge of nonmonotonic agents.

In the next sections we will develop a series of logics that, starting from the basic reasoning machinery of the system (represented by the logic $\mathcal{L}_0$) will first include introspection (the logic $\mathcal{L}_1$), and then empathy (the logic $\mathcal{L}_2$). There is a developmental metaphor to the sequence of $\mathcal{L}_0$, $\mathcal{L}_1$, and $\mathcal{L}_2$. $\mathcal{L}_0$ represents the reasoning ability of a system reasoning about the world (in a first-order language). The logic $\mathcal{L}_1$ represents the system when it "learns" to introspect, to refer to its own knowledge. Finally, $\mathcal{L}_2$ represents a further developmental stage, where the system realizes that other agents are just like itself, and learns to use empathy to reason about them.

Before we proceed it is important to point out that the logics that we will develop here are different from "standard" approaches to multiagent belief reasoning in two aspects. The first difference is that we will not really develop logics in the sense of describing either a proof theory or a model semantics for them. It turns out that the principles behind the idea of empathy will be formalized as constraints on the derivation relation (either proof theoretical derivation or semantic entailment) of these logics. We will still use the term "develop a logic" but the reader must be aware that we are referring to an implicit characterization of the logic through the constraints it must satisfy, instead of an explicit characterization.

The second difference with standard logics for multiagent belief reasoning is that we will use internal logics. An internal logic assumes an agent's point of view, whereas an external logic assumes a "reality's" point of view [McA88, Lev90]. The difference between an internal and an external logic refers to the question of how to interpret the assertion of a formula $\alpha$. In an internal logic, asserting $\alpha$ states that the particular agent whose point of view the logic assumes knows or believes $\alpha$. In a external logic, asserting $\alpha$ states that $\alpha$ is true, that it holds in the reality. Thus, in an internal logic, which is the approach we will take in this paper, formulas should be interpreted in relation to a particular agent's

[1] We will also use the neologism "empathic" to denote the corresponding adjective.

point of view. We will call this agent **the system** or **the reasoner**.

# 2 The logic of self-knowledge ($\mathcal{L}_1$)

## 2.1 The logic $\mathcal{L}_0$

Let us use the symbol $\mathcal{L}_0$ to represent the logic that models the system's basic, first-order reasoning machinery. We will use the symbol $\vdash_0$ to represent the derivation relation in $\mathcal{L}_0$. That is,

$$\alpha \vdash_0 \beta$$

states that given $\alpha$, the system would be able to derive $\beta$, where both $\alpha$ and $\beta$ are first-order formulas. Or in other words, if the reasoner believes $\alpha$ it would also believe $\beta$. We will also use the constants $\mathbf{0}$ and $\mathbf{1}$ to denote the reasoner and the other agent, respectively.

## 2.2 The logic $\mathcal{L}_1$

Syntactically, the logic $\mathcal{L}_1$ extends the logic $\mathcal{L}_0$ by including the modal operator $\mathbf{B}_0$ which represents the reasoner's self-knowledge, that is, knowledge about its own knowledge.[2] The derivation relation for $\mathcal{L}_1$ will be denoted by $\vdash_1$.

We will not allow for quantifying into a modal scope. That is, $\mathcal{L}_1$ will include formulas like $\mathbf{B}_0\neg\mathbf{B}_0(\exists x car(x) \wedge own(\mathbf{0}, x))$, but not formulas like $\mathbf{B}_0\neg\exists x(\mathbf{B}_0 car(x) \wedge own(\mathbf{0}, x))$.

## 2.3 Inclusion

The first constraint on the logic $\mathcal{L}_1$ is that it should include all the basic inference abilities of the reasoner. Using the developmental metaphor mentioned in the introduction, when the system learns to introspect it does not forget how to reason about the outside world. This can be captured formally as:

**Inclusion 1** $\quad$ if $\overline{\alpha} \vdash_0 \overline{\beta}$ then $\alpha \vdash_1 \beta$ $\quad$ (1)

which basically states that if something is provable in $\mathcal{L}_0$, it should remain provable in $\mathcal{L}_1$.

The operator $\overline{\alpha}$ is a "de-modalization" operator, a syntactic operation that uniformly replaces all modal subformulas in $\alpha$ by new propositional symbols. This is defined as:

$$\begin{aligned}
\overline{p} &= p \quad \text{if } p \text{ is a propositional symbol} \\
\overline{\alpha \wedge \beta} &= \overline{\alpha} \wedge \overline{\beta} \\
\overline{\alpha \vee \beta} &= \overline{\alpha} \vee \overline{\beta} \\
\overline{\alpha \rightarrow \beta} &= \overline{\alpha} \rightarrow \overline{\beta} \\
\overline{\neg\alpha} &= \neg\overline{\alpha} \\
\overline{\mathbf{B}_0\alpha} &= q \quad \text{where } q \text{ is a new propositional symbol}
\end{aligned}$$

[2] The name "auto-epistemic" knowledge would be more appropriate than "self-knowledge" but it would be too confusing with auto-epistemic nonmonotonic logics [Moo83, MT91].

The use of the “de-modalization” operator is necessary because the relation $\vdash_0$ is not be defined for modal formulas, and thus the expression $\alpha \vdash_0 \beta$ would be undefined if $\alpha$ or $\beta$ were modal formulas. This is of particular importance if $\vdash_0$ is the derivation relation of a non-monotonic logic that is not defined for formulas containing a belief operator, for example conditional logics [Del88, Bou92], or circumscription [McC80, McC86].

## 2.4 The self-knowledge inference rules

We will assume that the reasoner has positive introspection, and thus if it believes $\alpha$, it can conclude $\mathbf{B}_0\alpha$. This is captured by the following inference rule:

$$\textbf{Self-knowledge} \qquad \alpha \vdash_1 \mathbf{B}_0\alpha \tag{2}$$

We will also assume that the reasoner has privileged access to its knowledge, that is, if it believe that it believe $\alpha$ then it indeed believes $\alpha$. This is captured by:

$$\textbf{Privileged access} \qquad \mathbf{B}_0\alpha \vdash_1 \alpha \tag{3}$$

## 2.5 $\mathcal{L}_1$ is consistent

We would like to compare the logic $\mathcal{L}_1$ with some other, more conventional, modal logic for belief. But one has to remember that the conventional modal logics are usually understood as external logics, and thus in order to compare $\mathcal{L}_1$ with another modal logic, one has to “externalize” $\mathcal{L}_1$.

We will show that the logic $\mathcal{L}_1$ is consistent by showing that when one understands the logic $\mathcal{L}_1$ from an external point of view, it is contained in a logic that we called $qKT'4$ (after quantified, axiom K, a weaker version of the axiom T, and axiom 4), which in turn is weaker than the usual logics of knowledge, say KD45 [HM92].

**Theorem 1** *If $\mathcal{L}_0$ is sound and complete in relation to first-order logic, then if $\alpha \vdash_1 \beta$ then $\vdash_X \mathbf{B}_0\alpha \rightarrow \mathbf{B}_0\beta$, where $X$ a logic we will call $qKT'4$, and it is defined by the following axioms and inference rules:*

- $\models_X \alpha$ *if $\alpha$ is a first-order tautology*
- $\models_X \mathbf{B}_0\alpha \wedge \mathbf{B}_0(\alpha \rightarrow \beta) \rightarrow \mathbf{B}_0\beta$ *(the K axiom)*
- $\models_X \mathbf{B}_0\mathbf{B}_0\alpha \rightarrow \mathbf{B}_0\alpha$ *(the $T'$ axiom, which is a weaker version of the standard T axiom.)*
- $\models_X \mathbf{B}_0\alpha \rightarrow \mathbf{B}_0\mathbf{B}_0\alpha$ *(the 4 axiom)*
- $\dfrac{\alpha \quad \alpha \rightarrow \beta}{\beta}$ *Modus ponens*
- $\dfrac{\models_X \alpha}{\mathbf{B}_0\alpha}$ *Necessitation*

The theorem states that if the system that implements the logic $\mathcal{L}_1$ is able to derive $\beta$ from $\alpha$, then that could be described as a theorem of the logic $qKT'4$ that believing $\alpha$ implies believing $\beta$. Thus the logic $qKT'4$ can be seen as a logic that describes, from an external point of view, the behavior of a system that implements the logic $\mathcal{L}_1$.

We believe, but we have not been able to prove it yet, that $\mathcal{L}_1$ is "complete" in relation to $qKT'4$, that is if $\models_X \mathbf{B}_0\alpha \rightarrow \mathbf{B}_0\beta$ then $\alpha \vdash_1 \beta$. If that is true then $qKT'4$ is the external characterization of $\mathcal{L}_1$.

# 3 The logic of empathy ($\mathcal{L}_2$)

We will now define the logic $\mathcal{L}_2$, the logic of empathic reasoning. Syntactically, $\mathcal{L}_2$ extends $\mathcal{L}_1$ by including the modal operator $\mathbf{B}_1$ to denote the belief of the other agent. We will use the symbol $\vdash_2$ to denote the derivation relation in $\mathcal{L}_2$.

In terms of inference power, $\mathcal{L}_2$ must include $\mathcal{L}_1$. Formally:

**Inclusion 2** if $\alpha \vdash_1 \beta$ then $\alpha \vdash_2 \beta$

This states that if a certain derivation is possible in the logic of self-knowledge, then this derivation is also possible in $\mathcal{L}_2$. In fact, this constraint should also contain a syntactic operator that corresponds to the de-modalization operator of Inclusion 1, and that uniformly replaces subformulas that contain the modal operator $\mathbf{B}_1$ with new propositional symbols. For the sake of brevity we will leave the formal definition out.

## 3.1 Empathy

As we mentioned above, the idea of empathic reasoning about knowledge is to reason about the knowledge of an agent by disregarding our own beliefs and assuming the agent's beliefs and performing the inferences using those beliefs.

To define empathy we will define a substitution operation $|^{0,1}_{1,0}$ which substitutes in parallel $\mathbf{B}_0$ by $\mathbf{B}_1$, $\mathbf{B}_1$ by $\mathbf{B}_0$, $\mathbf{0}$ by $\mathbf{1}$, and $\mathbf{1}$ by $\mathbf{0}$. Formally:

- if $\alpha$ is a first order formula then $\alpha|^{0,1}_{1,0}$ is the same formula $\alpha$ where all $\mathbf{1}$ has been substituted by $\mathbf{0}$ and all $\mathbf{0}$ has been substituted by $\mathbf{1}$ in parallel.
- $(\mathbf{B}_0\alpha)|^{0,1}_{1,0} \quad = \quad \mathbf{B}_1(\alpha|^{0,1}_{1,0})$
- $(\mathbf{B}_1\alpha)|^{0,1}_{1,0} \quad = \quad \mathbf{B}_0(\alpha|^{0,1}_{1,0})$

The empathy principle is expressed as the following constraint on the $\vdash_2$ relation:

**Empathy** $\alpha \vdash_2 \beta$ if and only if $\mathbf{B}_1(\alpha|^{0,1}_{1,0}) \vdash_2 \mathbf{B}_1(\beta|^{0,1}_{1,0})$ (4)

The empathy principle can be interpreted in a goal-based way. If the system needs to know what the other agent will conclude given that (the system believes that) the other agent believes $\alpha$ ($\mathbf{B}_1\alpha$), it will apply the substitution operator to $\alpha$ (since the operation is its own inverse), perform its own reasoning using $\alpha|_{1,0}^{0,1}$ as the premise, and whatever conclusions it derives (say $\beta$), transfer it to the other agent by applying the substitution operation $(\mathbf{B}_1\beta|_{1,0}^{0,1})$.

Thus, the reasoner can deduce that the other agent has positive introspection, that is, if it believes *alpha* then it believes that it believes *alpha*, by the following steps:

1. from $\mathbf{B}_1 alpha$
2. apply the substitution operator to *alpha*, $\alpha|_{1,0}^{0,1}$
3. perform its own inference process in $\alpha|_{1,0}^{0,1}$. In this case the reasoner performs positive introspection (in fact a deduction possible in $\vdash_1$, which is included in $\vdash_2$) and deduces that $\mathbf{B}_0\alpha|_{1,0}^{0,1}$
4. apply the substitution operator to that result: $(\mathbf{B}_0\alpha|_{1,0}^{0,1})|_{1,0}^{0,1} \quad \equiv \quad \mathbf{B}_1\alpha$
5. and transfer that formula to the other agent's knowledge space: $\mathbf{B}_1\mathbf{B}_1\alpha$
6. and thus $\mathbf{B}_1\alpha \vdash_2$
   $bel_1\mathbf{B}_1\alpha$.

### 3.2 $\mathcal{L}_2$ is consistent

We can prove that $\mathcal{L}_2$ is consistent by showing that it is included in a logic similar to $qKT'4$, but with two modal operators. We call this logic 2-$qKT'4$.

**Theorem 2** *If $\mathcal{L}_0$ is sound and complete in relation to first-order logic then, if $\alpha \vdash_2 \beta$ then $\models_Y \mathbf{B}_0\alpha \rightarrow \mathbf{B}_0\beta$, where $Y$ is the modal logic 2-$qKT'4$, defined as $qKT'4$ for both modal operators.*

We also conjecture that $\mathcal{L}_2$ is "complete" in relation to 2-$qkT'4$, that is, if $\models_Y \mathbf{B}_0\alpha \rightarrow \mathbf{B}_0\beta$, then $\alpha \vdash_2 \beta$. If this is true then 2-$qKT'4$ is the external characterization of $\mathcal{L}_2$. Even without the prove of the completeness in relation to 2-$qKT'4$, we are able to prove that $\mathcal{L}_2$ is capable of interesting forms of reasoning, for example a form of generalized empathy: if $Q$ is any sequence of operators $\mathbf{B}_0$ and $\mathbf{B}_1$, and $\alpha$ and $\beta$ are first-order formulas, then if $\alpha \vdash_0 \beta$ then $Q\alpha \vdash_2 Q\beta$. This, for example, would account for the "whales are not mammals" example in this paper's introduction.

## 4 Extensions to empathy

### 4.1 Lack of knowledge

The empathy logic as presented here does not deal with lack of knowledge ($\neg\mathbf{B}$). This is exactly the source of differences of this logic from more "standard" logics for belief, say KD45 or K45, and why we have to characterize the logic $\mathcal{L}_2$ using more "exotic" logics like 2-$qKT'4$. The main problem with dealing with lack of knowledge is that it cannot be represented using an internal logic. There seems to be no way of capturing, for example, negative introspection ($\neg\mathbf{B}\alpha \rightarrow \mathbf{B}\neg\mathbf{B}\alpha$ in an external logic), since there is no way to represent the system's lack of knowledge. Of course one can represent other agent's lack of knowledge, but that would not be derived through empathic means from the system's own self-knowledge.

On the other hand, it is possible to include a weaker version of the consistency axiom (D) in the logic $\mathcal{L}_1$, which would be propagated to $\mathcal{L}_2$. This weaker version of D, which we will call D$'$, can be included as a property of the $\vdash_1$ relation.

$$\mathbf{D}' \quad \alpha \vdash_1 \neg\mathbf{B}_0\neg\alpha$$

D$'$ corresponds, from an external point of view, to the axiom:

$$\mathbf{B}_i\alpha \rightarrow \mathbf{B}_i\neg\mathbf{B}_i\neg\alpha$$

## 4.2 Multi-agent extension

At the moment the logic $\mathcal{L}_2$ deals with only two agents: the reasoner and "the other." This can be extended to many agents. The logic $\mathcal{L}_3$ extends $\mathcal{L}_2$ by including a finite set of operators $\mathbf{B}_i$ that represents the beliefs of agent $i$. The empathy rule is exactly the same as $\mathcal{L}_2$ with the same motivation: if the system is able to perform a reasoning, anybody else can do it too. The only difference is in the substitution operation, which now will be denoted by $|^{0,i}_{i,0}$. This substitution operator is similar to $|^{0,1}_{1,0}$, and is defined as:

- if $\alpha$ is a first order formula then $\alpha|^{0,i}_{i,0}$ is the same formula $\alpha$ where all $\mathbf{i}$ has been substituted by $\mathbf{0}$ and all $\mathbf{0}$ has been substituted by $\mathbf{i}$ in parallel.
- $(\mathbf{B}_0\alpha)|^{0,i}_{i,0} \quad = \quad \mathbf{B}_i(\alpha|^{0,i}_{i,0})$
- $(\mathbf{B}_i\alpha)|^{0,i}_{i,0} \quad = \quad \mathbf{B}_0(\alpha|^{0,i}_{i,0})$
- $(\mathbf{B}_j\alpha)|^{0,i}_{i,0} \quad = \quad \mathbf{B}_j\alpha \quad \text{if } i \neq j$

Empathy in $\mathcal{L}_3$ is expressed as:

$$\textbf{Empathy 2} \qquad \alpha \vdash_3 \beta \quad \text{if and only if} \quad \mathbf{B}_i(\alpha|^{0,i}_{i,0}) \vdash_3 \mathbf{B}_i(\beta|^{0,i}_{i,0}) \tag{5}$$

## 4.3 Reasoning about other modalities

We can also extend the empathic principle to reason about other modalities in addition to belief. If the logic $\mathcal{L}_0$ is not a first-order logic, but say a modal logic that models also goals and actions, then the resulting $\mathcal{L}_2$ would, in a limited way, extend it to two or more agents setting. If the logic $\mathcal{L}_0$ includes say a modal operator for goals $\mathbf{G}_0$, we will need to include in the corresponding $\mathcal{L}_2$ logic the modalities $\mathbf{G}_0$ and $\mathbf{G}_1$, and redefine the substitution and de-modularization operators accordingly. The resulting $\mathcal{L}_2$ would correctly deal with formulas of the form $\mathbf{B}_1\mathbf{G}_0\alpha$, for $\alpha$ first-order. But the resulting logic would not deal with formulas where a belief operator is within the scope of a goal operator, for example $\mathbf{G}_1\mathbf{B}_0\alpha$.

### 4.4 Non-monotonic empathy

Although we have been dealing with monotonic reasoning so far, the empathic form of reasoning can be naturally extended to non-monotonic reasoning. If the basic logic $\mathcal{L}_0$ is nonmonotonic, then the resulting $\mathcal{L}_3$ empathic logic will also be nonmonotonic. The constraints on $\mathcal{L}_1$ and $\mathcal{L}_2$ (and the principles behind them) are invariant whether $\mathcal{L}_0$ is monotonic or not. In fact, this seems to be the great strength of the principle.

Let us call the basic nonmonotonic logic as n$\mathcal{L}_0$, and the logics derived from them as n$\mathcal{L}_1$ and n$\mathcal{L}_2$. n$\mathcal{L}_2$ has some of the properties that [Wai93a] claim are "good" properties of a nonmonotonic logic to reason also about nonmonotonic agents. The principle of Empathy extends the *internal default* principle in that paper to deal with nested beliefs, and the principle of Inclusion 1, due to its de-modalization operator, seems to be the correct characterization of the principle *external default*. On the other hand, n$\mathcal{L}_2$ will not exhibit the *epistemic cancellation* property described there, since it involves lack of knowledge.

## 5 Related Results and Conclusions

This work is an extension of the one reported in [MWC91] where a weaker version of the empathy principle, named "projection," was introduced. A similar work [Wai93b] explores the basic idea of the empathy principle within an external logic framework.

The idea of empathy is certainly not new; different versions of it have appeared, usually implicitly, in many papers. It is probably one of those ideas that are so obvious that no one can claim its authorship. Neither would the author. What this paper hopes to have accomplished is a formalization of such idea. From this point of view, this work can be best compared with [Kon86]. The main difference is that [Kon86] assumes that the inference machinery of the reasoner is sound and complete with relation to some of the standard modal logics for belief (KD45 and others). In this work we make no such assumptions; we "build" the reasoner's inference machinery based on its original first-order machinery plus the principles of inclusion 1 and 2, the principles of self-knowledge and privileged access, and the principle of empathy. We prove that the resulting logic is at least sound with respect to the logic 2-$qKT'4$ (or any stronger modal logic).

Furthermore we believe that the formalization of the empathic reasoning formalized here can serve as a stepping stone for more complex forms of reasoning about other agents, like reasoning about other modalities and nonmonotonic reasoning. One can develop a single agent nonmonotonic logic, and the correspondent empathic extension would allow the logic to deal with many agents. Of course the problem is not totally solved, since the empathic extension of a logic $\mathcal{L}_0$ is only implicitly defined, which may not give any help on determining a semantics or a proof theory for the extended logic.

## References

[Bou92] Craig Boutilier. Conditional logics for default reasoning and belief revision. Technical Report KRR–TR–92–1, University of Toronto, Computer Science Department, 1992.

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

[HM92] J. Y. Halpern and Y. Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54(3):319–379, 1992.

[Kon86] K. Konolige. *A Deduction Model of Belief.* Morgan Kaufman, Los Altos, CA, 1986.

[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.

[McC86] John McCarthy. Applications of circumscription to formalizing common-sense reasoning. *Artificial Intelligence*, 28:89–116, 1986.

[Moo83] R. C. Moore. Semantical considerations on nonmonotonic logic. In *IJCAI-83*, pages 272–279, Karlsruhe, West Germany, 1983.

[MT91] W. Marek and M. Truszczyński. Autoepistemic logic. *Journal of the ACM*, 38:588–619, 1991.

[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.

[Wai93a] Jacques Wainer. Epistemic extension of preference logics. In *Proc. of the 13th International Joint Conference on Artificial Intelligence*, pages 382–387, 1993.

[Wai93b] Jacques Wainer. Introspection and projection in reasoning about other agents. In *Anais do X Simpósio Brasileiro de Inteligência Artificial (SBIA 93)*, Porto Alegre, RS, Outubro 1993.