Introspection and Reasoning about the Beliefs of other Agents[1]

Anthony S. Maida

Department of Computer Science
The Pennsylvania State University

ABSTRACT

A cognitive agent uses representations to reason about the world. An "introspective" cognitive agent has the ability to manipulate representations (meta representations) of its own representations. If such an agent were to embody its beliefs in its representations, then the agent could reason about its own beliefs by manipulating its meta representations.

A "belief reasoner" can reason about the beliefs of other agents. There has been considerable research in the construction of belief reasoners. This paper observes that the construction of such systems can, in large part, be reduced to the task of constructing introspective systems.

We illustrate how an introspective agent can use analogy-based reasoning to construct an architecture for belief reasoning on the basis of examining its own architecture.

1. Introduction.

This paper views the task of reasoning about the beliefs of other agents in terms of more basic processes of introspection in which an agent reasons about its own beliefs. We begin with the assumption that agents think in a proposition-like "mentalese." The intuition underlying our line of argument is as follows: If agents think in a proposition-like mentalese, then to describe their belief states, we can use a metalanguage capable of describing arrangements of propositions. If the language is capable of describing any possible arrangement of propositions, then the language is capable of describing any possible belief state of the agent. If an agent already uses such a language to introspectively describe and reason about its own beliefs, it should be possible for this agent to adapt this language, by some analogy-based process, to describe the beliefs of other agents.

2. Cognitive Science: The Representational Theory of Mind

The notion that an agent's beliefs determine its behavior is compatible with the metaphor of a "knowledge-based system" as used in artificial intelligence or the "representational theory and computational theory of mind" as used by philosophers of cognitive science (e.g., Fodor, 1981). In this view, a cognitive agent represents the world by the use of some internal language, a "mentalese," such as a propositional language. This view has led to the "knowledge representation

[1] This research was supported by ONR contract N00014-85-K-0521. Thanks to Ross Canfield, Minqui Deng, Minkoo Kim, Drew McDermott, Joe Niederberger, and Bonnie Webber for help with various incarnations of this manuscript.

hypothesis," upon which much of artificial intelligence and cognitive science is based (cf, Fodor, 1980; Pylyshyn, 1984; Smith, 1982). Thus, the primary assumption of this paper is compatible with much of mainstream cognitive science.

2.1. The Semantics of Belief Sentences.

To say that an agent has a belief is to say that the agent has constructed a representation in its mental language and that the agent takes this representation as accurately describing something. For instance, say that an agent named "Pat" believes that a dog named "Fido" is ferocious. This means that in the mental language which Pat uses to represent the world, there is an expression that resides in his data base, which represents the proposition *Fido is ferocious*. It also means that Pat bears some relation to this expression indicating that it is believed by Pat.

If we assume that Pat thinks with propositions, then the sentence "Pat believes Fido is ferocious" can be taken to be true exactly if the proposition (1):

(1) (is-ferocious Fido)

resides in Pat's data base and the proposition is somehow marked as true (possibly implicitly, by virtue of it simply residing in the knowledge base). We have sketched for the above belief sentence a semantics in terms of the propositional content of a knowledge base and we will call this a *knowledge based semantics*.[2]

2.2. Belief Spaces as Nested Mental Models.

A perspicuous notation for depicting the beliefs of cognitive agents is by the use of belief spaces (cf., Fauconnier, 1985). A belief space can be construed as a mental model of another agent's representation of the world.[3] This naturally leads to an architecture which is a tree of nested mental models or belief spaces (cf., Fauconnier, 1985; Maida, 1984). Figure 1 depicts such a tree.

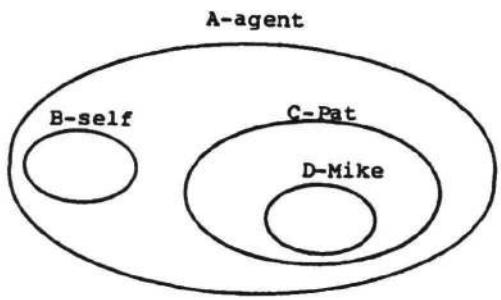

Figure 1. A tree of belief spaces. Nested ellipses indicate subtrees.

Each ellipse in this figure is a belief space; the nestings indicate nestings of the belief spaces. The ellipse labeled "A" indicates the agent's knowledge (beliefs). The ellipse labeled "B" indicates the agent's knowledge about its own knowledge; that is, objects in this space represent objects in the parent space. Ellipse "C" indicates the agent's knowledge about Pat's knowledge; that is, objects in this space represent objects in Pat's data base (not shown). Finally, the system has knowledge that Pat has knowledge of Mike's knowledge (indicated with ellipse D); that is, objects in this space represent objects in Pat's space that represent objects in Mike's space.

---

[2] This contrasts with the possible-world semantics of Hintikka (1962). See Halpern & Moses (1985) for a guide to modal logics of knowledge and belief. The possible-worlds approach attempts to define knowledge without reference to the internal structure of the agent who has the knowledge. The approach described in section 2.1 is sometimes called the "syntactic approach."

[3] Belief spaces or analogs have been used by Moore (1973), Cohen (1978), Martins & Shapiro, 1983; Rapaport & Shapiro, 1984; and Kobsa (1985). In linguistics, Fauconnier (1985) has made heavy use of belief spaces.

Simulative Reasoning. *Simulative reasoning*[4] is the process of one agent reasoning about the beliefs of a second agent as if the second agent were reasoning with his own beliefs. The belief-space architecture is highly suggestive of simulative reasoning. In principle, one could have inference processes in each belief space. In this paper, we will assume there is one actual inference engine in the root space. To conduct simulative reasoning in a child space, the inference engine will simulate a virtual inference engine in the child space.

2.3. Introspection as Representations of Representations.

Suppose we have a cognitive agent who not only maintains representations which describe things external to itself, but also maintains representations of its representations. The agent could then have beliefs about its representations and we could say that the agent is introspective.

3. Reducing Belief Reasoning to Introspection.

To some, the belief-space architecture may seem unparsimonious. Would it be plausible to assume that a cognitive agent would just happen to have an architecture consisting of a tree of data bases, simply to reason about others' beliefs? Fortunately, we can reduce the belief-space architecture to more basic principles. The belief-space architecture seems to be a consequence of any knowledge-based system that has a sufficiently rich self-model and capacity to reason by analogy.

The basic idea is to have the system replicate a theory of its own inference ability in a model of another agent, so that the copy is suitably modified to appropriately describe that other agent. We will call such an operation *projective analogy.*[5]

3.1. Replication of One's Self-Model and Inference Machinery.

Suppose an agent has a partial description of its own structure, including a descriptive sketch of the operations of its own analogy-based reasoning ability. This is illustrated in Figure 2a below. The large circle indicates the set of propositions that the agent believes. The inner circle, labeled "self-model," consists of the set of propositions that describe the agent's beliefs. Note that this "self-model" is the agent's self image. For reasoning about its beliefs, the agent has access only to the information delimited by the inner circle (i.e., its self-model), and not the outer circle. If the agent were to make an analogy between its own structure and the structure of some other entity, the analogy would have to be based on its self-model.

If the system learned that another agent, say Pat, has structure similar to it, then the system could create a description of Pat, shown in Figure 2b, by making a copy of its own self-model and modifying it appropriately to apply to Pat. (For instance, this might be done by replacing all occurrences of the symbol "self" in the description with the symbol "Pat.")

There is one more step needed in order to construct a tree of belief spaces. Suppose the agent's self-model contains a description of the analogy process. If so, then a copy of this description will have been duplicated in the description for Pat. If the agent then reasons about what would happen to Pat's data base if he (reciprocally) thought about the system, it will conclude that Pat will construct, within his own data base, a model of the system. We will call this *reciprocal* projective analogy. That is, *the original agent can apprehend that Pat can use his own reasoning to reason about it.* Hence, we have sketched how the construction of a tree of nested belief spaces might be automated. The next section will describe an example.

---

[4] Creary (1979) seems to have been the first to use the term. Dinsmore (1985) discusses simulative reasoning from a linguistic perspective.

[5] We call this "projective analogy" because, first, it is a kind of analogy, and second, the process of attributing one's own characteristics to another is called "projection" by some psychologists.

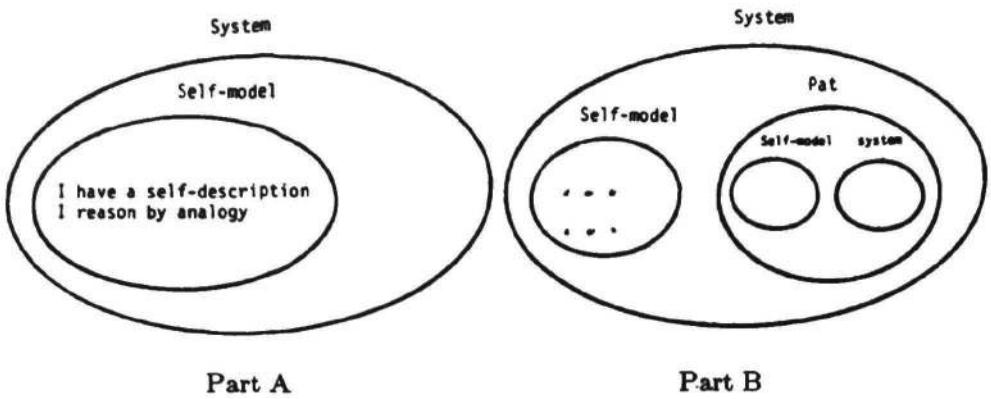

Figure 2. If a system with a self-model (A) learns of an entity, say Pat, who has similar structure to itself, then the system can create a model of Pat by projective analogy (B).

4. A Detailed Example.

In this section, we give a detailed example of the projective analogy process. It is the simplest example we can think of. We are going to get an agent to realize that another agent (Pat) can reason by *modus ponens*. However, we will fail to get the agent to realize that Pat realizes that it (the original agent) reasons by *modus ponens*. The reason for the failure will be that the original agent does not have an explicit model of the analogy process; the agent does not realize that it reasons by analogy and thus cannot attribute this characteristic to another Pat.

For our example we will break down the process of evolving from an introspective agent into a belief reasoner into two steps. They are:

> We must give the agent a theory of its own inference ability. We will call this an *auto-rational theory.*[6]
>
> We also need a capacity to replicate this theory and modify the copy to describe the inference ability of another agent; that is, a capacity for *projective analogy.*

The auto-rational theory may only partially describe the agent's inference ability. In our case, the auto-rational theory will not describe the capacity for projective analogy. With a partial theory, the replicated-and-modified theory constitutes a partial theory of the other agent.

4.1. A Partial Auto-Rational Theory.

We will postulate an agent that has three rules of inference wired into its mental structure. These are: 1) a procedural version of *modus ponens;* 2) an ability to do *propositional introspection;* and, 3) an ability for *projective analogy.* These three abilities characterize our agent's capacity to do inference. If the agent can represent explicitly that it has these abilities, then it has a theory of its own rationality--an auto-rational theory.

Modus ponens and *propositional introspection* are defined below. The letters p and q range over propositional formulas and the symbol *bt* stands for "believes that."

> **Modus Ponens.** If the propositions p and $p->q$ reside in the agent's data base, then the proposition q will reside in the agent's data base.
>
> **Propositional Introspection.** If the proposition p resides in the agent's data base, then the proposition *(bt self "p")* resides in the agent's data base.[7]

---

[6] Moore (1985) used the term *auto-epistemic* to refer to a theory of one's own knowledge. We are concerned with a particular aspect of one's own knowledge, namely one's knowledge of his rationality. An auto-rational theory is a kind of auto-epistemic theory.

[7] We will assume that a proposition is marked as true if it resides in the data base.

Suppose further that the axiom schemas (a)-(d) below reside in the agent's data base.

(a) (bt self "(p & p->q) -> q)")
(b) (bt self "(bt self "p") -> (bt self "(bt self "p")")")
(c) ((bt self "p") & (bt self "p->q")) -> (bt self "q")
(d) (bt self "p") -> (bt self "(bt self "p")")

Axioms (a)-(d) present a partial auto-rational theory of the agent's reasoning. Expressions (a) and (b) are *descriptive* but they are not *causal*. Axiom (a) is true exactly if the agent reasons by *modus ponens* as the axiom describes. However, the agent's believing the axiom does not cause the agent to reason by *modus ponens*. Axiom (b) is true exactly if the agent can do *propositional introspection*.

Expressions (c) and (d) are causal because they can be interpreted directly by the inference engine. They enable the system to do simulative reasoning about its beliefs in its self-model. We will call expression (a) the assumption of *awareness of rationality;* (b) the assumption of *awareness of propositional introspection;* (c) the assumption of *auto-syllogistic interpretation;* and (d) the assumption of *introspective interpretation*.

Auto-syllogistic interpretation allows the agent to simulate reasoning by *modus ponens* in its self-model. *Introspective interpretation* allows an agent to simulate *propositional introspection* in its self-model.

4.2. Genesis of Attributions of Rationality in Others.

Figure 3a, using the belief-space notation, depicts an agent with the above auto-rational theory. Expressions (a) and (b) are in the agent's self-model. Expressions (c) and (d) are in the root space. Figure 3b depicts the agent after the theory has been replicated and modified by projective analogy to describe another agent, Pat. The replicated expressions are indicated by apostrophes.

Figure 3. Replication of a partial auto-rational theory.

We will now describe the projective analogy process. It involves two components: 1) universal generalization;[8] and 2) universal quantifier elimination. *Universal generalization* transforms a copy of expression (a) into (q) by replacing a constant (which must be of type "agent") with a universally quantified (actually quantified over agents) variable as shown below:

(a) (bt self "((p & p->q) -> q)")

⇓ Universal Generalization on Agents

(q) (forall (x) (bt x "((p & p->q) -> q))")

[8] Universal generalization is a form of inductive reasoning. In this paper it will be the means by which we make a generalization about agents.

Universal quantifier elimination transforms an expression such as (q) into (a') below by instantiating the variable with a constant.

(a') (bt Pat "((p & p->q) -> q)")

Upon applying the composition of universal quantifier elimination and universal generalization to each of the expressions (a), (b), (c), and (d) we get expressions (a'), (b'), (c'), and (d') below.

(a') (bt Pat "((p & p->q) -> q)")
(b') (bt Pat "(bt Pat "p") -> (bt Pat "(bt Pat "p")")")
(c') ((bt Pat "p") & (bt Pat "p->q")) -> (bt Pat "q")
(d') (bt Pat "p") -> (bt Pat "(bt Pat "p")")

The expressions (a')-(d') describe the replicated and modified structure of Figure 3b. We started with an agent who had the mental representation of Figure 3a, and by projective analogy, the agent arrived at the structure of Figure 3b. Expressions (a') and (b') comprise the agent's model of Pat. Expressions (c') and (d') allow the agent to do simulative reasoning in Pat's belief space.

The agent can attribute to Pat only what it realizes about itself. Notice that the agent does not attribute an ability to do projective analogy to Pat. This is because the agent does not explicitly that it itself has this ability. Notice also that the agent has a model of itself, but does not view Pat has having a model of himself.

4.3. Awareness that One has a Self-Model.

Without a self-model, the agent would not realize that it thinks. However, to realize that it has a self-model, the agent must have a model of its self-model. It is for this reason that the projective analogy process did not attribute a self-model to Pat. Based on analogy with itself, the system did not realize that it had a self-model. However, the system has the ability to infer a model of its initial model. It can do simulative reasoning of itself in its self-model.

Self-Replication of One's Self-Model. As the agent thinks with its self-model, this can cause it to realize that it has a self-model. The agent can actually construct a copy of its self-model in its self-model. This is done as follows.

Since the agent can reason by *modus ponens*, it can manipulate expressions (a) and (d) to get expression (aa) below. Similarly, the agent can manipulate expressions (b) and (d) to get (bb) below. Since the agent can reason by *propositional introspection*, it can apply this to expression (c) to get expression (cc) below. The agent can also apply *propositional introspection* to get expression (dd) below.

(aa) (bt self "(bt self "(p & p->q) -> q)")")
(bb) (bt self "(bt self "(bt self "p") -> (bt self "(bt self "p")")")")
(cc) (bt self "((bt self "p") & (bt self "p->q")) -> (bt self "q")")
(dd) (bt self "(bt self "p") -> (bt self "(bt self "p")")")

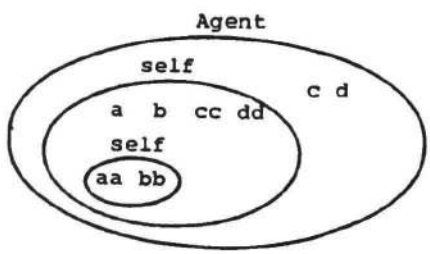

Figure 4. Using simulative reasoning to create a model of one's self-model.

4.4. Simulative Reasoning.

With what we have so far, the agent can do simulative reasoning in Pat's subspace by virtue of expression (d'). Since the agent can reason by *modus ponens*, it can manipulate expressions (a') and (d') to get expression (aa') below. Similarly, the agent can manipulate expressions (b') and (d') to get (bb') below.

(aa') (bt Pat "(bt Pat "(p & p->q) -> q")")

(bb') (bt Pat "(bt Pat "(bt Pat "p") -> (bt Pat "(bt Pat "p")")")")

Expressions (aa') and (bb') can be construed as Pat's self-model within the system's model of Pat, as shown in Figure 5.

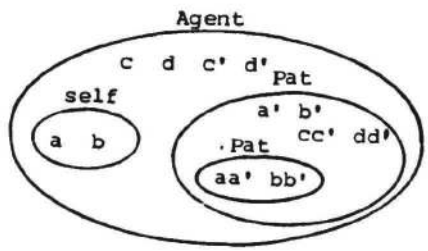

Figure 5. Using simulative reasoning to create a model of another's self-model.

4.5. Reciprocal Projective Analogy.

In section 4.4 we saw a replicated-and-modified model of Pat (i.e., expressions (a') and (b')) partially undergo another replicate-modify cycle producing expressions (aa') and (bb'). However, this new model is not very functional. For instance, there are no formulas analogous to (dd') which would enable simulative reasoning in that space. Additionally, this second-generation model cannot even partially replicate.

Realizing that Others Believe that You are Rational. For the agent to believe that Pat could realize that the agent itself is rational, we would need to create a model of the agent within Pat's subspace. We have called this *reciprocal projective analogy*. This would require simulative reasoning in Pat's subspace, using the rule of projective analogy. This cannot be done because the agent does not model Pat as having the ability to do projective analogy.

5. Summary and Conclusions.

So far we have achieved the following. We have shown how it might be possible for one agent to infer that a second agent has beliefs and that this second agent realizes it itself has beliefs. We have argued theoretically that it should also be possible to infer that this second agent can infer that other agents have beliefs, and that those other agents can make similar inferences. In essence this amounts to an architecture of belief spaces. The importance of this is that the intuitive notation of belief spaces can be interpreted as a kind of cognitive architecture with possible psychological reality. The architecture would have a natural explanation.

Limitations and Further Work. The language we have been using to express autorational theories has been, for the most part, *propositional* as opposed to predicate based. That is, it views propositions as atomic and cannot describe their subparts. This means that the language is inherently unable to describe various interesting phenomena. In particular, *projective analogy* cannot be described in the language because that would involve the reference to subparts of propositions. Consequently, we cannot have a fully introspective agent who reasons by analogy if his declarative mental language is only propositional. It would never be able to represent to itself the process of *projective analogy*. The reason we could not get our agent to realize that Pat realized it was rational was traced back to the fact that our original agent did not have an explicit model of the projective analogy process. It appears that this cannot be remedied in a propositional framework.

The robustness of the projective analogy process we described should also be scrutinized. Our characterization is really a kludge to demonstrate the feasability of the idea. It is unlikely to be robust. The domain of projecting properties from oneself to others should simply be another domain to study analogical reasoning.

In summary, topics for future research include the following: 1) we need more detailed introspective models, particularly beyond the propositional level; and, 2) we need to look at more cases of an agent reasoning by analogy from his own structure to make inferences about another agent.

References

[1] Cohen, P. R. On knowing what to way: Planning speech acts. Ph.D. Thesis, Technical Report No. 118, Department of Computer Science, University of Toronto, January, 1978.

[2] Creary, L. G. Propositional attitudes: Fregean representation and simulative reasoning. *Proc. IJCAI, 1979, 6, 176-181.*

[3] Dinsmore, J. Mental Spaces from a Functional Perspective. Manuscript, Department of Computer Science, Southern Illinois University at Carbondale, 1985.

[4] Doyle, J. A Model for Deliberation, Action, and Introspection. Doctoral dissertation submitted to the Massachusetts Institute of Technology; also M.I.T. Artificial Intelligence Lab. Memo AIM-TR-581, 1980.

[5] Fauconnier, G. *Mental spaces: Aspects of meaning construction in natural language.* Cambridge: MIT Press, 1985.

[6] Fodor, G. Methodological solipsism considered as a research strategy in cognitive psychology. In John Haugeland (Ed.), *Mind Design* Cambridge: MIT Press, 1981.

[7] Halpern J. Y. & Moses Y. A Guide To The Modal Logics Of Knowledge And Belief: Preliminary Draft. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence,* Los Angeles, California, August 18-23, 1985, Vol. 1, pp. 480-490.

[8] Hintikka, J. *Knowledge and Belief.* Cornell University Press, Ithaca, New York, 1962.

[9] Kobsa, A. Using situation descriptions and Russellian attitudes for representing beliefs and wants. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence,* Los Angeles, California, August 18-23, 1985, Vol. 1, pp. 513-515.

[10] Maida, A. S. Selecting a humanly understandable representation for reasoning about knowledge. *International Journal of Man-Machine Studies,* 22, 1985, 151-161.

[11] Maida, A. S. & Kim, M. Belief, Equality, and Quantification in the Belief Space Engine. Unpublished manuscript, 1985a.

[12] Maida, A. S. & Kim, M. The Belief Space Engine user's manual (draft). Department of Computer Science, Penn State University, University Park, PA, 16802, August, 1985b.

[13] Martins, J. P. & Shapiro, S. C. Reasoning in multiple belief spaces. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence,* Karlsruhe, West Germany, August 8-12, 1983, pp. 371-373.

[14] Moore R. C. D-SCRIPT: A computational theory of definite descriptions. *Advance papers from the Third International Joint Conference on Artificial Intelligence,* Stanford, California, August, 1973, pp. 223-229.

[15] Moore, R. C. Reasoning about knowledge and action. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence,* Cambridge, Massachusetts, August, 1977, pp. 223-227.

[16] Moore, R. C. Semantical considerations in nonmonotonic logic. *Artificial Intelligence,* 25(1), 1985, 75-94.

[17] Pylyshyn, Z. W. *Computation and cognition: toward a foundation for cognitive science.* Cambridge: MIT Press, 1984.

[18] Rapaport, W. J. & Shapiro, S. C. Quasi-indexical reference in propositional semantic network. *Proceedings of the Tenth International Conference on Computational Linguistics,* Stanford, California, July 2-6, 1984, pp. 65-70.

[19] Smith, B. C. Reflection and Semantics in a Procedural Language. Ph. D. Thesis and Tech. Report MIT/LCS/TR-272, MIT, Cambridge, MA, 1982.