

Maximal Introspection of Agents

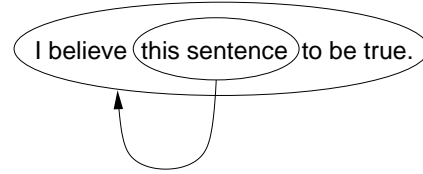
Thomas Bolander
Technical University of Denmark

Main goal: to avoid the paradoxes of self-reference and introspection when formalizing theories for reasoning about multi-agent systems (using the syntactical treatment).

1

Introspective beliefs are beliefs that an agent has of its own beliefs.

Examples. “I do not believe that everything I believe about computer science is true”. Can even be **self-referential**:



More seriously, in multi-agent systems self-reference can be **indirect**:

Agent 1: Everything agent 2 believe, I believe.

↓

Agent 2: Everything agent 1 believe, I believe.

↑

2

Two Treatments of Belief

To reason about agents we formalize their beliefs as formulas in formal theories. We have two choices of formalization:

- **Semantical treatment:** treating belief as an *operator*:

$$B_i\varphi \text{ or } \Box\varphi$$

using *modal logic*.

- **Syntactical treatment:** treating belief as a *predicate*:

$$B_i(\ulcorner\varphi\urcorner), \text{ where } \ulcorner\cdot\urcorner \text{ is some coding,}$$

using *first-order predicate logic*.

3

- Advantages of syntactical treatment:

- **More expressive:** e.g. “agent 1 has no contradictory beliefs” formalized by

$$\forall x (\neg B_1(x) \vee \neg B_1(\text{not } x))$$

has no semantical counterpart ($\Box x$ is not well-formed).

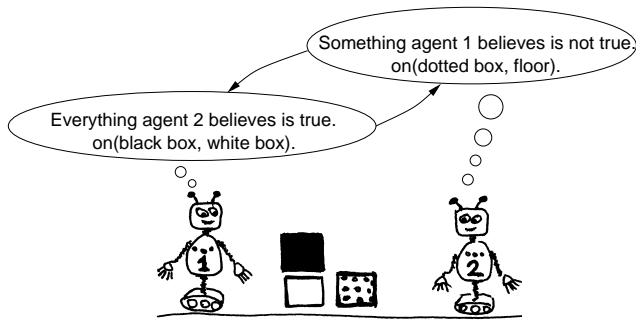
- No logical omniscience: beliefs not preserved under logical equivalence.
- “Directly” implementable in LP or a first-order theorem prover.

- Disadvantage of syntactical treatment:

- Representing beliefs syntactically easily leads to inconsistency of representing system. This is because agents can have (indirect) self-referential beliefs which produces paradoxes within the representing system.

4

Example—Syntactical Treatment



Assume everything agent 2 believes is true: then, in particular, something agent 1 believes is false. But, by assumption, both of agent 1's beliefs are true. This is a contradiction.

Assume something agent 2 believes is false: then it must be false that agent 1 has a false belief. If all beliefs of agent 1 are true then, in particular, everything agent 2 believes is true. Again, a contradiction.

5

This is a *paradox*.

Conclusion: Any reasoning framework in which this is a possible scenario will be inconsistent.

Based on this conclusion, our *main problem* becomes:

to suitably restrict the syntactic treatment such that consistency of the representing systems can always be guaranteed.

—//—

Note: Notion of *truth* in example above can be replaced by *belief* (but argument becomes more complex).

6

Restriction Strategies

Different strategies to avoid the paradoxes (inconsistency):

- (i) **Stratification of language.**
Disadvantage: we can not quantify over *all* beliefs. E.g. $\forall x (\neg B_1(x) \vee \neg B_1(\text{not } x))$ becomes impossible.
- (ii) **Weakening of underlying logic** (e.g. 3-valued, paraconsistent).
Disadvantage: accept sentences not concerning multi-agent system.
- (iii) **Restricting language of agents**, i.e. restrict the set of formulas φ for which we allow $B_i(\ulcorner \varphi \urcorner)$ to occur in our theory.
Advantage: Only exclude sentences producing problems and keep everything else (more fine-grained than (i) and (ii)).

We consider here (iii).

7

The Main Result

Def. A belief $B_i(\ulcorner \varphi \urcorner)$ is said to be **negatively quantified** if some $B_j(x)$ occurs quantified in the scope of \neg . E.g. $B_2(\ulcorner \exists x (B_1(x) \rightarrow \neg B_2(x)) \urcorner)$.

Theorem 4. If we avoid negatively quantified beliefs then our reasoning framework can not become inconsistent.

Thus it is always safe for agents to express:

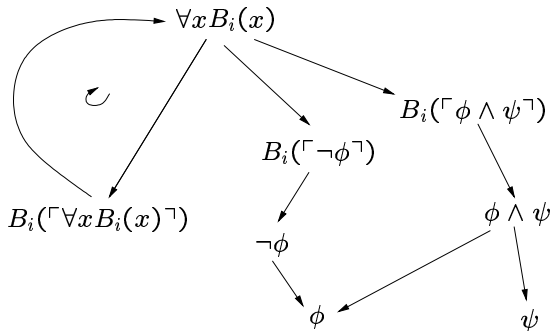
- *Nested beliefs* as e.g.
 $B_2(\ulcorner B_1(\ulcorner \text{on}(\text{black box, floor}) \urcorner) \urcorner)$.
- *Negated beliefs* as e.g.
 $B_2(\ulcorner \neg B_1(\ulcorner \text{on}(\text{dotted box, floor}) \urcorner) \urcorner)$
(generalizes the main result of Perlis [1986]).
- *Quantified beliefs* as e.g.
 $B_1(\ulcorner \forall x (Q(x) \rightarrow B_1(x) \vee B_2(x)) \urcorner)$
(generalizes the main result of des Rivières & Levesque [1988]).

8

Method of Proof

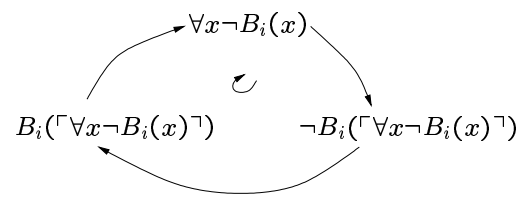
If *self-reference* is the problem, we first need to make the notion precise.

Formula $B_i(\ulcorner \varphi \urcorner)$ can be thought of as belief *referring* to φ . $\varphi \wedge \psi$ can be thought of as (semantically) *referring* to both φ and ψ . References can be collected into a **reference graph**:



Def. A sentence is **self-referential** if it is contained in a cycle in the graph.

9



To obtain consistency self-reference must either be *neutralized* or *excluded*.

Proving consistency amounts to proving existence of truth-value assignment to nodes of graph (using fixed point results for chain complete partial orders (ccpo's)).

Now the Main Result (Theorem 4) follows from:

- Disallowing quantified beliefs \Rightarrow no cycles in graph \Rightarrow self-reference is *excluded*.
- Disallowing negated beliefs \Rightarrow no negation in cycles \Rightarrow self-reference is *neutralized*.

10

Conclusion

- The syntactical treatment of belief is *prima facie* unattractive because (indirectly) self-referential beliefs cause paradoxes and inconsistency.
- There is a cure: restricting the language of the agents such as to disallow them expressing viciously self-referential beliefs.
- We are thereby able keep the advantages of the syntactic treatment over the semantic treatment—in particular its higher expressive power allowing agents to have general beliefs about the beliefs of other agents.
- We have provided a general method for proving consistency of such restricted frameworks: fixed points on reference graphs.

11