Electronic Notes in Theoretical Computer Science 70 No. 5 (2002)
URL: http://www.elsevier.nl/locate/entcs/volume70.html 16 pages

Maximal Introspection of Agents

Thomas Bolander [1]

Informatics and Mathematical Modelling
Technical University of Denmark
Copenhagen, Denmark

Abstract

This paper concerns the representation of introspective belief and knowledge in multi-agent systems. An introspective agent is an agent that has the ability to refer to itself and reason about its own beliefs. It is well-known that representing introspective beliefs is theoretically very problematic. An agent which is given strong introspective abilities is most likely to have inconsistent beliefs, since it can use introspection to express self-referential beliefs that are paradoxical in the same way as the classical paradoxes of self-reference. In multi-agent systems these paradoxical beliefs can even be expressed as beliefs about the correctness and completeness of other agents' beliefs, i.e., even without the presence of explicit introspection. In this paper we explore the maximal sets of introspective beliefs that an agent can consistently obtain and retain when situated in a dynamic environment, and when treating beliefs "syntactically" (that is, formalizing beliefs as axioms of first-order predicate logic rather than using modal formalisms). We generalize some previous results by Perlis [1985] and des Rivières & Levesque [1988].

1 Introduction

Formal languages and theories can be used to represent and reason about agents and their beliefs about the world (including other agents and the agents themselves). A large number of different types of languages and theories have been proposed to this aim. Theories of first-order predicate logic seem particularly attractive for this purpose because of their high expressive power, and because of the extensive use of first-order logic [2] in computational systems such as the ones used in logic programming. But first-order theories have two major drawbacks:

(i) *Complexity*: provability in first-order logic is only semi-decidable.

[1] Email: tb@imm.dtu.dk

[2] By "first-order logic" we everywhere mean "first-order predicate logic".

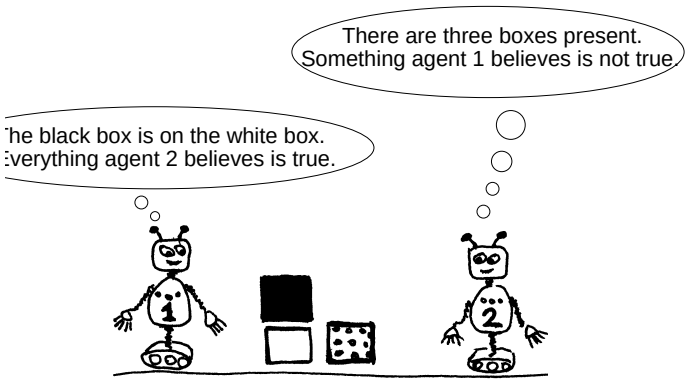

Fig. 1. Two agents having beliefs about each other.

(ii) *Inconsistency*: representing *introspective* beliefs of agents (that is, the kind of beliefs that agents have of their own beliefs) often leads to paradoxes and inconsistency of the systems in which these beliefs are represented.

In this paper we will concentrate on ways to circumvent (ii) such that we can have consistent first-order logics of introspective beliefs—that is, we concentrate on ways to obtain consistent formalisms for introspection while still retaining the expressiveness of first-order logic.

The paper is organized as follows. In Section 2 we give an informal motivating example showing how indirect introspective beliefs can lead to inconsistency. In Section 3 we introduce more formally the "syntactic approach" to belief representation, and introduce a number of general principles of belief (called *epistemic principles*) which we expect our agents to satisfy. In Section 4 it is shown, though, that in general we cannot make our agents satisfy these principles. More precisely, it is shown that theories containing these principles will be inconsistent. In Section 5 we consider ways of restricting the principles such that we can have consistent theories containing them. The main result in this direction is Theorem 5.9. It is shown that there does not exist consistent principles much stronger than those proven to be consistent by this theorem.

2 A Motivating Example

Consider the situation depicted in Figure 1. We have here a blocks world with two agents, agent 1 (on the left) and agent 2 (on the right). We assume that the beliefs of agent 1 are given completely by the following two propositions S_1 and T_1:

S_1 : The black box is on the white box.
T_1 : Everything agent 2 believes is true.

Furthermore, we assume the beliefs of agent 2 to be given completely by:

S_2 : There are three boxes present.
T_2 : Something agent 1 believes is not true.

The reason for agent 2 to believe that agent 1 has a false belief could e.g. be that agent 2 sees that agent 1 cannot see the dotted box (the black box is blocking the view of agent 1), and agent 2 therefore expects agent 1 to have the false belief that there are only two boxes present.

The problem with the presented two-agent situation is that it is *paradoxical*. It turns out that no matter whether we assume T_1 to be a true or a false proposition, we are lead to a contradiction. The argument is as follows:

Assume first that T_1 is true: then everything agent 2 believes is true, and in particular T_2 must is true. That is, something agent 1 believes is not true. But this is a contradiction, since S_1 is definitely true, and T_1 is true by assumption.

Assume now that T_1 is false: then something agent 2 believes must be false. But since S_2 is true, it must be T_2 that is false. From this it follows that every belief of agent 1 must be true. But T_1 is among agent 1's beliefs, and this proposition is assumed to be false. Again we have a contradiction.

That is, it is neither consistent to assume that T_1 is true, nor that it is false. This is a paradox. The conclusion we have to draw from this paradox is that any formal framework for reasoning about agents in which the situation in Figure 1 is possible must be inconsistent.

At first sight the example does not seem to have to do with introspection and self-reference, since the agents have no explicitly given beliefs about themselves. But introspection and self-reference is obtained indirectly: the belief T_1 of agent 1 refers to every belief of agent 2. In particular T_1 refers to T_2, which in turn refers back to T_1. It is this presence of indirect self-reference in the beliefs of the agents that leads to the paradox.

The argument of the paradox given above can be formalized in first-order logic, thus showing that not all beliefs can be treated consistently when formalized in first-order logic. Theorem 4.2 in Section 4 gives a number of examples of beliefs of agents that makes the first-order theory in which they are formalized inconsistent.

Based on the paradox and its formalizability in first-order logic, our main goal becomes:

> *to find suitable restrictions on the (indirect) introspective and self-referential beliefs such that consistency can be ensured.*

This is is the goal that we will pursue throughout this article.

3 Representing Beliefs of Agents

Representing beliefs of agents as axioms of first-order predicate logic is called the *syntactic* approach to belief representation. In the syntactic approach, a belief is represented as a formula $B_i(\ulcorner\varphi\urcorner)$ where B_i is a predicate symbol, φ is a sentence of first-order logic (possibly containing B_i), and $\ulcorner\cdot\urcorner$ is some coding scheme. By a **coding scheme** we understand any injective map $\ulcorner\cdot\urcorner$ from the set of sentences of the language in question into the set of closed terms of that language. $B_i(\ulcorner\varphi\urcorner)$ should be read as: "agent i believes that φ".

In the so-called *semantic* approach to belief representation, B_i is a modal operator rather than a predicate symbol. To express that "agent i believes φ" we would then simply write $B_i\varphi$. In the semantic approach no coding is needed.

The syntactic approach is preferred to the semantic one because of its expressiveness. In the syntactic approach a statement like e.g. "agent 1 has no contradictory beliefs" can be expressed by the formula $\neg\exists x(B_1(x) \wedge B_1(\mathit{not}(x)))$, assuming that we have appropriate axioms for the function *not*. But this statement can not be expressed in the semantic approach, since even if we have a modal logic with variables (such as first-order modal logic), the modal operator can not be applied directly to the variables—that is, $B_i x$ (or $\Box x$) is not a well-formed modal formula and therefore neither is $\neg\exists x\,(B_i x \vee \neg B_i x)$. The propositions T_1 and T_2 of the example above can not be expressed as sentences in a model logic either.

Through the use of axioms such as $B_i(\ulcorner\varphi\urcorner)$ we can construct formal theories to represent facts of, and to reason about, multi-agent systems. Such systems could for instance be distributed computer systems, where each process is considered to be an agent. It could also be e.g. systems of autonomous robots acting in some shared environment.

Example 3.1 Consider again the situation depicted in Figure 1. One way of representing this situation could be by a theory T including axioms

$$\mathit{on}(\mathit{white\ box}, \mathit{floor})$$

$$\mathit{on}(\mathit{black\ box}, \mathit{white\ box})$$

$$\mathit{on}(\mathit{dotted\ box}, \mathit{floor})$$

$$B_1\,(\ulcorner \mathit{on}(\mathit{white\ box}, \mathit{floor})\urcorner)$$

$$B_1\,(\ulcorner \mathit{on}(\mathit{black\ box}, \mathit{white\ box})\urcorner)$$

$$B_2\,(\ulcorner \mathit{on}(\mathit{dotted\ box}, \mathit{floor})\urcorner)$$

$$B_2\,(\ulcorner \exists x\,(\mathit{is\text{-}box}(x) \wedge \mathit{on}(\mathit{black\ box}, x))\urcorner)$$

$$B_i\,(\ulcorner \mathit{on}(\mathit{black\ box}, \mathit{white\ box}) \rightarrow \neg\mathit{on}(\mathit{black\ box}, \mathit{floor})\urcorner) \quad \text{for } i = 1, 2$$

where $B_1\,(\ulcorner \mathit{on}(\mathit{white\ box}, \mathit{floor})\urcorner)$ means that agent 1 believes the white box to

be on the floor, and $B_2\,(\ulcorner \exists x(\textit{is-box}(x) \wedge \textit{on}(\textit{black box}, x))\urcorner)$ means that agent 2 believes that the black box is on some other box (but he cannot see which one, since the dotted box is blocking his view).

When having described a multi-agent system as a formal theory T, reasoning about this system and the beliefs of its agents amounts to proving theorems in T, since every theorem of T becomes a fact concerning the system.

The theory T given above is quite weak and does not allow us to deduce much about the beliefs of agent 1 and agent 2. We even cannot infer

$$B_i(\ulcorner \neg \textit{on}(\textit{black box}, \textit{floor})\urcorner)$$

from

$$B_i\,(\ulcorner \textit{on}(\textit{black box}, \textit{white box}) \rightarrow \neg \textit{on}(\textit{black box}, \textit{floor})\urcorner)$$

and

$$B_i(\ulcorner \textit{on}(\textit{black box}, \textit{white box})\urcorner),$$

since we have no axioms or inference rules allowing this. Therefore, to make such theories useful as reasoning mechanisms, we should include general epistemic principles such as e.g.

$$B_i(\ulcorner \varphi \rightarrow \psi\urcorner) \rightarrow (B_i(\ulcorner \varphi\urcorner) \rightarrow B_i(\ulcorner \psi\urcorner)) \quad \text{for all } i \text{ and all sentences } \varphi, \psi$$

meaning that the beliefs of all agents are closed under modus ponens.

Below we give a list of some of the most common such principles, translated into first-order logic from the corresponding principles in modal logic. These principles are, for all sentences φ, ψ in our language,

(R2)	$\dfrac{\varphi}{B_i(\ulcorner \varphi\urcorner)}$
(A1)	$B_i(\ulcorner \gamma\urcorner)$ when γ is a valid sentence in first-order predicate logic
(A2)	$B_i(\ulcorner \varphi \rightarrow \psi\urcorner) \rightarrow (B_i(\ulcorner \varphi\urcorner) \rightarrow B_i(\ulcorner \psi\urcorner))$
(A3)	$B_i(\ulcorner \varphi\urcorner) \rightarrow \varphi$
(A4)	$B_i(\ulcorner \varphi\urcorner) \rightarrow B_i\,(\ulcorner B_i(\ulcorner \varphi\urcorner)\urcorner)$
(A5)	$\neg B_i(\ulcorner \varphi\urcorner) \rightarrow B_i\,(\ulcorner \neg B_i(\ulcorner \varphi\urcorner)\urcorner)$
(D)	$B_i(\ulcorner \varphi\urcorner) \rightarrow \neg B_i(\ulcorner \neg\varphi\urcorner)$
(BC1)	$B_i\,(\ulcorner B_j(\ulcorner \varphi\urcorner) \rightarrow \varphi\urcorner)$
(BC2)	$B_i\,(\ulcorner \varphi \rightarrow B_j(\ulcorner \varphi\urcorner)\urcorner)$
(O)	$\varphi \leftrightarrow B_i(\ulcorner \varphi\urcorner)$

(BC1) says that agent i believes all of agent j's beliefs to be correct (true). When $i = 1$ and $j = 2$ this expresses the belief T_1 of agent 1 in the example of

Section 2. **(BC2)** says that agent i believes agent j to believe everything that is correct (true). **(O)** is called the **omniscience principle**. This is a very strong principle saying that everything believed is true, and that everything true is believed. Agents satisfying this principle are called **omniscient**: they possess complete and correct knowledge of the world. **(O)** is recognized to be identical to Tarski's schema for truth, schema T, in first-order logic (see e.g. [12]). We should not in general expect agents to satisfy a strong principle such as **(O)**, but we include it among our epistemic principles anyway since, as we will see in Section 5, proving consistency results for restricted versions of **(O)** will automatically give us consistency results for restricted versions of all the other epistemic principles.

In many cases we might not want our epistemic principles to be instantiated with every single sentence of the language. For instance, an agent might only be omniscient with respect to some small part of the world. If this part of the world can be described through the sentences of some set M, then we could have

$$\varphi \leftrightarrow B_i(\ulcorner \varphi \urcorner) \quad \text{for all } \varphi \in M$$

instead of

$$\varphi \leftrightarrow B_i(\ulcorner \varphi \urcorner) \quad \text{for } \textit{all} \text{ sentences } \varphi.$$

Let $\mathcal{L}$ denote any first-order language. Let M be a set of sentences of $\mathcal{L}$, that is, let $M \subseteq \mathcal{L}$.[3] By $\mathbf{(O)}_M$ we understand the set of instances of **(O)** with sentences of M. $\mathbf{(O)}_M$ is called the **omniscience principle** over M. $\mathbf{(R2)}_M$, $\mathbf{(A1)}_M$–$\mathbf{(A5)}_M$, $\mathbf{(D)}_M$, $\mathbf{(BC1)}_M$, and $\mathbf{(BC2)}_M$ are defined similarly to be the set of instances of the respective epistemic principles over M.

Definition 3.2 Let M be a set of sentences of some first-order language $\mathcal{L}$. By an **epistemic theory** over M we understand any combination of

$$\mathbf{(R2)}_M, \mathbf{(A1)}_M\text{–}\mathbf{(A5)}_M, \mathbf{(D)}_M, \mathbf{(BC1)}_M, \mathbf{(BC2)}_M, \text{and } \mathbf{(O)}_M,$$

that is, any combination of the epistemic principles given above with φ and ψ instantiated over M.

4 The Problem of Obtaining Consistency

We can think of the epistemic principles as something being added to a *base theory*. By a **base theory** we understand any theory T in $\mathcal{L}$ satisfying:[4]

(i) If $\varphi \in T$ then either φ does not contain any of the B_i's, or $\varphi = B_i(\ulcorner \psi \urcorner)$ for some ψ not containing any of the B_i's.

(ii) If $B_i(\ulcorner \varphi \urcorner) \in T$ then $T \vdash \varphi$.

These two characterizing properties can be paraphrased as:

[3] We identify first-order languages with their set of sentences.

[4] A theory T in $\mathcal{L}$ is simply a set of sentences of $\mathcal{L}$.

(i') There are no *meta-beliefs* (beliefs about beliefs) in T.

(ii') Everything believed by an agent in T is true (in T).

The theory considered in Example 3.1 is a base theory. A base theory describes the environment in which the agents are situated as well as the agents' first-order beliefs about this environment. Condition (ii) simply says that all (first-order) beliefs about the environment are correct. Agents might of course in some situations have false beliefs, but we do not consider such beliefs as part of the base theory.

When we construct a theory T for reasoning about a multi-agent system, our choice of epistemic principles should not depend on our choice of base theory. For one thing, the epistemic principles only give general properties of belief (or knowledge), and our choice of these should only depend on what kind of "modality" (or propositional attitude) B_i is supposed to capture. Another thing is that in a changing environment the base theory might change over time to reflect these changes in the environment, but this should not affect the epistemic principles and their validity. But, surprisingly, it turns out that whether a theory including a set of epistemic principles is consistent or not depends crucially on the chosen base theory. To see this, let us first introduce the notion of *universal consistency*.

Definition 4.1 An epistemic theory E is called **universally consistent** if, for any consistent base theory B, the theory $B \cup E$ is consistent.

All epistemic theories used to reason about agents should of course be universally consistent: if not, we could end up in that very peculiar situation that our theory could suddenly become inconsistent just from updating some axioms of the base theory. But:

Theorem 4.2 *No epistemic theory extending any of the following theories is universally consistent.*

(1) ***(R2) + (A1) + (A3)****.*

(2) ***(A1) + (A2) + (A3) + (BC1)****.*

(3) ***(A1) + (A2) + (A4) + (D) + (BC1)****.*

(4) ***(A1) + (A2) + (D) + (BC1) + (BC2)****.*

Proof. Trivially, when a theory is not universally consistent, neither is any extension of it. We therefore only need to prove that none of (1)–(4) is universally consistent. That (1) and (2) are not universally consistent is a direct consequence of a theorem of Montague in [7]. That (3) is not universally consistent is a direct consequence of a theorem of Thomason in [13]. Both these theorems are reviewed in [3] and [8]. Finally, that (4) is not universally consistent is proved in Section 5.3. □

The fact that (1) is not universally consistent is probably not a serious problem, since **(R2)** is a very strong principle that beliefs of agents would not

be likely to satisfy (at least not in the syntactic treatment). We might also in many cases take (2) to be too strong, since **(A3)** excludes the possibility of an agent having a false belief (though, in other cases, when reasoning about agents, it seems appropriate to assume that the agents will never believe anything that is not true). But, as we see from (3) and (4), even if **(R2)** and **(A3)** are excluded, we can still get an epistemic theory which is not universally consistent.

The theorem shows, as also suggested by the example of Section 2, that assuming agents to have certain seemingly natural beliefs and assuming their beliefs to satisfy certain seemingly natural principles can make the entire reasoning framework in which these beliefs are represented inconsistent.

Based on this negative result our main problem now becomes: *to find sets of sentences M with which we can safely instantiate our epistemic theories, that is, to find sets $M \subseteq \mathcal{L}$ for which the epistemic theories over M are universally consistent.* This problem is the subject of the following section.

5 Some Universally Consistent Epistemic Theories

In this section we let $\mathcal{L}$ denote a fixed first-order language. We assume that $\mathcal{L}$ contains a number of unary predicate symbols $B_1, B_2, \ldots, B_n$ $(n > 0)$. In the following we will, for simplicity, only concentrate on finding sets $M \subseteq \mathcal{L}$ for which $\mathbf{(O)}_M$ is universally consistent. We can do this without loss of generality, for, as the following lemma shows, if $\mathbf{(O)}_M$ is universally consistent then so is any other epistemic theory over M.

Lemma 5.1 *Let $M \subseteq \mathcal{L}$ be a set of sentences satisfying*

$$\text{if } \varphi, \psi \in M \text{ then } B_i(\ulcorner \varphi \urcorner), \neg\varphi, \varphi \rightarrow \psi \in M.^{5}$$

If the omniscience principle over M, $\boldsymbol{(O)}_M$, is universally consistent, then every epistemic theory over M is universally consistent.

Proof. Assume M satisfies the requirement, and that $\mathbf{(O)}_M$ is universally consistent. Let B be an arbitrary consistent base theory. We have to show that then

$$C = B \cup \mathbf{(R2)}_M \cup \mathbf{(A1)}_M \cup \cdots \cup \mathbf{(A5)}_M \cup \mathbf{(D)}_M \cup \mathbf{(BC1)}_M \cup \mathbf{(BC2)}_M$$

is also consistent. Since we know $\mathbf{(O)}_M$ to be universally consistent, $B \cup \mathbf{(O)}_M$ must be consistent. If we therefore can prove that

$$\mathbf{(O)}_M \vdash \mathbf{(R2)}_M \cup \mathbf{(A1)}_M \cup \cdots \cup \mathbf{(A5)}_M \cup \mathbf{(D)}_M \cup \mathbf{(BC1)}_M \cup \mathbf{(BC2)}_M$$

then C must be consistent as well, and this will conclude the proof. We only prove $\mathbf{(O)}_M \vdash \mathbf{(D)}_M$. The rest of the epistemic principles can in the same

[5] That is, M is closed under application of B_i, $\neg$, and $\rightarrow$.

way easily be shown to follow from $\mathbf{(O)}_M$. To show that $\mathbf{(O)}_M \vdash \mathbf{(D)}_M$, let φ be any sentence in M. Then we have the following proof in $\mathbf{(O)}_M$:

1.	$\varphi \leftrightarrow B_i(\ulcorner\varphi\urcorner)$	axiom
2.	$\neg\varphi \leftrightarrow B_i(\ulcorner\neg\varphi\urcorner)$	axiom (M is closed under $\neg$)
3.	$\varphi \leftrightarrow \neg B_i(\ulcorner\neg\varphi\urcorner)$	by 2.
4.	$B_i(\ulcorner\varphi\urcorner) \leftrightarrow \neg B_i(\ulcorner\neg\varphi\urcorner)$	by 1. and 3.
5.	$B_i(\ulcorner\varphi\urcorner) \rightarrow \neg B_i(\ulcorner\neg\varphi\urcorner)$	by 4.

showing that the φ-instance of **(D)**, $B_i(\ulcorner\varphi\urcorner) \rightarrow \neg B_i(\ulcorner\neg\varphi\urcorner)$, holds in $\mathbf{(O)}_M$.□

The above lemma tells us that if we know $\mathbf{(O)}_M$ to be universally consistent then *any* epistemic theory can safely be instantiated with each of the sentences of M.

5.1 Well-founded Epistemic Theories

Our first universal consistency result is a slightly generalized version of a result of des Rivières & Levesque [3]. We start out with a couple of new definitions. A coding scheme $\ulcorner\cdot\urcorner$ is said to be **well-founded** if there is no infinite sequence of sentences $\varphi_0, \varphi_1, \varphi_2, \ldots$ such that for all $i \in \mathbb{N}$, φ_i contains $\ulcorner\varphi_{i+1}\urcorner$ as a term. In the following we will assume all considered coding schemes to be well-founded. [6]

Definition 5.2 Let φ, ψ be formulas of $\mathcal{L}$. We say that φ is **contained in** ψ if one of the following is the case

(i) φ is a sub-formula of ψ

(ii) there is a sub-formula $B_i(\ulcorner\gamma\urcorner)$ of ψ such that φ is contained in γ.

Note, that "contained in" above is defined recursively, and the definition only makes sense when the coding scheme is well-founded. By our definition, if φ and ψ are formulas and A is a one-place predicate symbol different from all the B_i's, then φ is contained in e.g. both $\varphi \wedge \psi$ and $B_i\,(\ulcorner B_i(\ulcorner\varphi\urcorner)\urcorner)$ but not in $A(\ulcorner\varphi\urcorner)$.

The following theorem tells us that we can always safely instantiate our epistemic theories with sentences such as

$$B_1\,(\ulcorner B_2\,(\ulcorner\neg on(\mathit{black\ box}, \mathit{floor})\urcorner)\urcorner)$$

[6] It can easily be seen that in any first-order language containing infinitely many ground terms it is possible to construct a well-founded coding scheme. Furthermore, any "standard" Gödel coding will obviously be well-founded.

in which we have nested beliefs, but not necessarily with sentences such as

$$B_1\,(\ulcorner \exists x(B_1(x) \wedge \neg B_2(x))\urcorner)$$

in which we have quantified beliefs.

Theorem 5.3 (After des Rivières & Levesque [3]) *The theory*

$$\{\varphi \leftrightarrow B_i(\ulcorner \varphi \urcorner) \mid \varphi \text{ is a sentence not containing } B_i(x)\}$$

is universally consistent. Thus, by Lemma 5.1, any epistemic theory over a set of sentences not containing $B_i(x)$ is universally consistent.

Example 5.4 The theorem shows that consistency is ensured if we refrain from expressing quantified beliefs. In the example of Section 2 both T_1 and T_2 are quantified beliefs, quantifying over the beliefs of agent 2 and agent 1, respectively. The theorem tells us that we would not be able to derive a paradox if the agents had no such quantified beliefs. Without quantified beliefs we can always reason about multi-agent systems consistently.

We will prove the theorem by another method than that of des Rivières & Levesque, who used a careful translation from a modal logic to prove their result. The important thing about our method is that it can be considered as a general method for proving these kinds of universal consistency results.

For every set $M \subseteq \mathcal{L}$ we define a functional $F_M : (\mathcal{L} \to \{t, f\}_\perp) \to (\mathcal{L} \to \{t, f\}_\perp)$ [7] by, for all sentences $\varphi, \psi, \forall x\gamma(x) \in \mathcal{L}$,

$$F_M([\![\cdot]\!])(\varphi \wedge \psi) = \begin{cases} t & \text{if } [\![\varphi]\!] = [\![\psi]\!] = t \\ \perp & \text{if } [\![\varphi]\!] = \perp \text{ or } [\![\psi]\!] = \perp \\ f & \text{otherwise} \end{cases}$$

$$F_M([\![\cdot]\!])(\neg\varphi) = \begin{cases} f & \text{if } [\![\varphi]\!] = \text{t} \\ t & \text{if } [\![\varphi]\!] = \text{f} \\ \perp & \text{otherwise} \end{cases}$$

$$F_M([\![\cdot]\!])(\forall x\gamma(x)) = \begin{cases} t & \text{if } [\![\gamma(\tau)]\!] = t \text{ for all terms } \tau \text{ in } \mathcal{L} \\ \perp & \text{if } [\![\gamma(\tau)]\!] = \perp \text{ for some term } \tau \text{ in } \mathcal{L} \\ f & \text{otherwise} \end{cases}$$

$$F_M([\![\cdot]\!])(B_i(\ulcorner \varphi \urcorner)) = [\![\varphi]\!] \text{ for all } \varphi \in M$$

The first three conditions above are recognized to correspond to Kleene's weak three-valued logic with $\perp$ as the third value. The fourth condition will ensure

[7] $\{t, f\}_\perp$ denotes the set $\{t, f, \perp\}$. t denotes "true" and f denotes "false". Mappings $\mathcal{L} \to \{t, f\}_\perp$ are used to represent *partial* functions, where the value $\perp$ means "undefined" (for further information see e.g. [11]).

that in any fixed point [8] of F_M we will have

$$[\![B_i(\ulcorner\varphi\urcorner)]\!] = F_M([\![\cdot]\!])(B_i(\ulcorner\varphi\urcorner)) = [\![\varphi]\!] \text{ for all } \varphi \in M. \quad (1)$$

Let $F : (\mathcal{L} \to \{t, f\}_\perp) \to (\mathcal{L} \to \{t, f\}_\perp)$ be a functional. The set of **initial elements** of F, denoted $\text{init}(F)$, is defined as

$$\text{init}(F) = \{\varphi \in \mathcal{L} \mid F([\![\cdot]\!])(\varphi) = \perp \text{ for all } [\![\cdot]\!] : \mathcal{L} \to \{t, f\}_\perp\}.$$

For F_M we have

$$\text{init}(F_M) = \{\varphi \in \text{atoms}(\mathcal{L}) \mid \varphi \text{ is not on the form } B_i(\ulcorner\psi\urcorner) \text{ for some } \psi \in M\}.$$ [9]

By an **initial extension** of F we understand any functional G that extends [10] F and satisfies

$$G([\![\cdot]\!])(\varphi) \in \{t, f\} \text{ for all } \varphi \in \text{init}(F) \text{ and all } [\![\cdot]\!] : \mathcal{L} \to \{t, f\}_\perp.$$

Now we can state an important lemma:

Lemma 5.5 *Let M be a subset of $\mathcal{L}$ containing at least all sentences in which none of the B_i's occur. If every initial extension of F_M has a total fixed point* [11] *then $(\mathbf{O})_M$ is universally consistent.*

Proof. [12] Assume every initial extension of F_M has a total fixed point. Let B be any consistent base theory. We need to prove that $B \cup (\mathbf{O})_M$ is consistent. Since B is consistent there exists some language $\mathcal{L}' \supseteq \mathcal{L}$ in which B has a Herbrand model H. Define an initial extension J of F_M by using the truth-values from H as values for the initial elements of F_M. By assumption, J has a total fixed point $[\![\cdot]\!]$. Let H' be the Herbrand interpretation [13] defined by

$$H' = \{\varphi \in \text{atoms}(\mathcal{L}') \mid [\![\varphi]\!] = t\}.$$

A simple induction proof on the syntactic complexity of φ now shows that for all $\varphi \in \mathcal{L}'$ we have

$$H' \models \varphi \Leftrightarrow [\![\varphi]\!] = t. \quad (2)$$

This is proved using the definition of H' and the first three defining conditions of F_M. Now, by (1) above we have $[\![B_i(\ulcorner\varphi\urcorner)]\!] = [\![\varphi]\!]$ for all $\varphi \in M$. Using (2),

[8] A function $[\![\cdot]\!] : \mathcal{L} \to \{t, f\}_\perp$ is a **fixed point** of a functional $G : (\mathcal{L} \to \{t, f\}_\perp) \to (\mathcal{L} \to \{t, f\}_\perp)$ if $G([\![\cdot]\!]) = [\![\cdot]\!]$.

[9] $\text{atoms}(\mathcal{L})$ denotes the set of ground atoms in $\mathcal{L}$.

[10] G **extends** F if for all maps $[\![\cdot]\!]$ and all φ, $F([\![\cdot]\!])(\varphi)$ is either undefined (has the value $\perp$) or has the same value as $G([\![\cdot]\!])(\varphi)$.

[11] A fixed point $[\![\cdot]\!] : \mathcal{L} \to \{t, f\}_\perp$ is total if it maps every $\varphi \in \mathcal{L}$ into $\{t, f\}$.

[12] The space available unfortunately only allow us to sketch the proofs. Contact the author to obtain the full proofs.

[13] As usual, Herbrand interpretations are identified with subsets of the Herbrand base (see e.g. [6]).

this gives us

$$H' \models \varphi \leftrightarrow B_i(\ulcorner \varphi \urcorner) \text{ for all } \varphi \in M.$$

This means that $\mathbf{(O)}_M$ holds in H'. Furthermore, from the choice of initial extension of F_M, it can be shown that B holds in H' as well. Thus H' is a model of $B \cup \mathbf{(O)}_M$, and this proves the consistency. □

We now define the notion of a *semantic graph*. For any functional F, we define the **semantic graph** of F as the directed graph $(\mathcal{L}, \rightarrow_F)$ with nodes $\mathcal{L}$, and edge relation $\rightarrow_F$ given by

$$\psi \rightarrow_F \varphi \quad \Leftrightarrow \quad \text{for all } [\![\cdot]\!], \text{ if } [\![\psi]\!] = \bot \text{ then } F([\![\cdot]\!])(\varphi) = \bot.$$

One way of expressing the condition on the right-hand side is that "the truth-value of ψ is needed to determine the truth-value of φ". In this sense, the semantic graph is a graph of *semantical dependency*: there is an edge from ψ to φ iff φ depends semantically on ψ.

Lemma 5.6 *Let $M \subseteq \mathcal{L}$. If the semantic graph of F_M is well-founded then any initial extension of F_M has a total fixed point.*

Proof. Let J be an initial extension of F_M. For $f, g : \mathcal{L} \rightarrow \{t, f\}_\bot$ we define $f \subseteq g$ to mean that everywhere f is defined, g has the same value as f (i.e., $\forall x(f(x) \neq \bot \rightarrow f(x) = g(x))$). It is easy to show that J is monotone on $\mathcal{L} \rightarrow \{t, f\}_\bot$ with respect to the ordering $\subseteq$. Furthermore, $\mathcal{L} \rightarrow \{t, f\}_\bot$ is a ccpo [14] with respect to $\subseteq$, and therefore J must have a least fixed point $[\![\cdot]\!]$ ([5]). To prove that $[\![\cdot]\!]$ is total, assume the opposite. Since the semantic graph $(\mathcal{L}, \rightarrow_J) = (\mathcal{L}, \rightarrow_{F_M})$ is well-founded, there must be a $\rightarrow_{F_M}$-minimal element on which $[\![\cdot]\!]$ is undefined. But this is easily seen to lead to a contradiction. □

Lemma 5.7 *The well-founded part of the semantic graph of $F_\mathcal{L}$ is the sub-graph induced by the set*

$$\{\varphi \in \mathcal{L} \mid \varphi \textit{ does not contain } B_i(x)\}.$$

We leave out the proof of this lemma. It is proven by considering the defining conditions of $F_\mathcal{L}$ and by using that $\ulcorner \cdot \urcorner$ is a well-founded coding scheme.

Now Lemma 5.5, 5.6, and 5.7 together immediately give us a proof of Theorem 5.3:

Proof. [of Theorem 5.3] Let $M = \{\varphi \in \mathcal{L} \mid \varphi \text{ does not contain } B_i(x)\}$. By Lemma 5.5 it is sufficient to prove that every initial extension of F_M has a total fixed point. But, by Lemma 5.7, the semantical graph of F_M is well-founded, and therefore Lemma 5.6 immediately gives the required result. □

[14] Chain complete partial order. See e.g. [5].

5.2 Positive Epistemic Theories

In the following we assume that the only propositional connectives used in first-order formulas are $\wedge$, $\vee$, and $\neg$ (i.e. not using $\rightarrow$ or $\leftrightarrow$).

The machinery we have introduced above also immediately gives a proof of the following theorem, which follows from a result proven independently by Perlis in [9] and Fefermann in [4].

Theorem 5.8 (After Perlis [9] and Fefermann [4]) *The theory*

$$\{\varphi \leftrightarrow B_i(\ulcorner\varphi\urcorner) \mid \varphi \text{ is a sentence of } \mathcal{L} \text{ in which no } B_i \text{ occurs in the scope of } \neg\}$$

is universally consistent. Thus, by Lemma 5.1, any epistemic theory over a set of sentences in which no B_i occurs in the scope of $\neg$ is universally consistent.

Proof. Let M be the set of positive sentences of $\mathcal{L}$. By Lemma 5.5 it is sufficient to prove that any initial extension of F_M has a total fixed point. Let J be any initial extension of F_M. Let $(\mathcal{L}, \rightarrow_J)$ be the semantic graph of J. Using Lemma 5.6 we can construct a fixed point on the well-founded part of the semantic graph. By Lemma 5.7, this gives us a fixed point $[\![\cdot]\!]'$ on the set of sentences not containing $B_i(x)$. Now define a functional G by

$$G([\![\cdot]\!])(\varphi) = [\![\varphi]\!]' \cup F_M([\![\cdot]\!])(\varphi) \ ^{15}$$

for all $[\![\cdot]\!]$ and φ. It can be seen that $G \restriction (\mathcal{L} \rightarrow \{t, f\})$ is monotone with respect to the ordering on $\{t, f\}$ given by $f < t$. Thus G has a total fixed point, and this will be a fixed point of F_M as well. □

5.3 Stronger Epistemic Theories

¿From the proof given for Theorem 5.8 above, we see that we have actually proven something even stronger:

Theorem 5.9 (Strengthening of Theorem 5.3 and Theorem 5.8) *The theory*

$$\{\varphi \leftrightarrow B_i(\ulcorner\varphi\urcorner) \mid \varphi \text{ is a sentence of } \mathcal{L} \text{ in which no } B_i(x) \text{ occurs in the scope of } \neg\}$$

is universally consistent. Thus, by Lemma 5.1, any epistemic theory over a set of sentences in which no $B_i(x)$ occurs in the scope of $\neg$ is universally consistent.

Example 5.10 This theorem gives us a larger set that we can safely instantiate our epistemic principles with than given by the two previous results. The

[15] For all $p, q \in \{t, f, \bot\}$, $p \cup q$ is the supremum of p and q wrt. to the ordering: $\bot < t, \bot < f$.

theorem proves that we can both safely instantiate with sentences such as

$$\neg B_i(\ulcorner on(black\ box, floor)\urcorner)$$

(which was not covered by Theorem 5.8) and sentences such as

$$\forall x\,(B_1(x) \wedge B_2(x))$$

(which was not covered by Theorem 5.3).

Let $M = \{\varphi \in \mathcal{L} \mid \text{no } B_i(x) \text{ occurs in the scope of } \neg \text{ in } \varphi\}$. It can easily be seen that M is very close to being a maximal set of sentences with which any epistemic principle can safely be instantiated. Actually, even a sentence such as $\forall x(A(x) \vee \neg B_i(x))$, which is one of the simplest sentences not in M, is not safe to instantiate with:

Lemma 5.11 *If M is a set containing the sentence $\forall x(A(x) \vee \neg B_i(x))$, where $A \neq B_i$, then $\boldsymbol{(O)}_M$ is not universally consistent.*

Proof. Let $\mathcal{L}$ be a first-order language with equality and let $\ulcorner\cdot\urcorner$ be any coding scheme in $\mathcal{L}$. Let $\varphi = \forall x\,(A(x) \vee \neg B_i(x))$ and let $M = \{\varphi\}$. We then have $\mathbf{(O)}_M = \{\varphi \leftrightarrow B_i(\ulcorner\varphi\urcorner)\}$. Let $B = \{\forall y\,(y = \ulcorner\varphi\urcorner \leftrightarrow \neg A(y))\}$. B is obviously consistent. We want to show that $B \cup \mathbf{(O)}_M$ is not, which proves that $\mathbf{(O)}_M$ is not universally consistent. To obtain a contradiction, assume that $B \cup \mathbf{(O)}_M$ is consistent. Then it has a model $\mathcal{M}$ in which $=$ denotes equality. This gives us the following sequence of implications

$$\begin{aligned}\mathcal{M} \models \varphi \Rightarrow \mathcal{M} \models \forall x\,(A(x) \vee \neg B_i(x)) \Rightarrow \mathcal{M} \models \forall x\,(x \neq \ulcorner\varphi\urcorner \vee \neg B_i(x)) \Rightarrow \\ \mathcal{M} \models \ulcorner\varphi\urcorner \neq \ulcorner\varphi\urcorner \vee \neg B_i(\ulcorner\varphi\urcorner) \Rightarrow \mathcal{M} \models \neg B_i(\ulcorner\varphi\urcorner) \Rightarrow \mathcal{M} \models \neg\varphi,\end{aligned}$$

which shows that $\mathcal{M} \not\models \varphi$. At the same time we have

$$\begin{aligned}\mathcal{M} \models \neg\varphi \Rightarrow \mathcal{M} \models \exists x\,(\neg A(x) \wedge B_i(x)) \Rightarrow \mathcal{M} \models \exists x\,(x = \ulcorner\varphi\urcorner \wedge B_i(x)) \Rightarrow \\ \mathcal{M} \models \ulcorner\varphi\urcorner = \ulcorner\varphi\urcorner \wedge B_i(\ulcorner\varphi\urcorner) \Rightarrow \mathcal{M} \models B_i(\ulcorner\varphi\urcorner) \Rightarrow \mathcal{M} \models \varphi,\end{aligned}$$

which shows that $\mathcal{M} \not\models \neg\varphi$. Now we have both $\mathcal{M} \not\models \varphi$ and $\mathcal{M} \not\models \neg\varphi$, which is a contradiction. □

This lemma also shows that (4) of Theorem 4.2 is not universally consistent. The argument is as follows. $\mathbf{(O)}$ is inconsistent by the above lemma, so for some sentence φ we have $\mathbf{(O)} \vdash \varphi$ and $\mathbf{(O)} \vdash \neg\varphi$. Now, for any theory U and any sentence φ, a simple induction on the proof length shows that

$$U \vdash \varphi \quad \Rightarrow \quad \{B_i(\ulcorner\varphi\urcorner) \mid \varphi \in U\} \cup \mathbf{(A1)} \cup \mathbf{(A2)} \vdash B_i(\ulcorner\varphi\urcorner).^{16} \qquad (3)$$

Letting $U = \mathbf{(O)}$, $\{B_i(\ulcorner\varphi\urcorner) \mid \varphi \in U\}$ becomes the theory $\mathbf{(BC1)} \cup \mathbf{(BC2)}$,

[16] We assume that first-order predicate logic is formulated such that the only rule of inference is modus ponens (see e.g. [1]).

and by (15) we then get

$$\textbf{(BC1)} \cup \textbf{(BC2)} \cup \textbf{(A1)} \cup \textbf{(A2)} \vdash B_i(\ulcorner \varphi \urcorner) \wedge B_i(\ulcorner \neg \varphi \urcorner)$$

which contradicts **(D)**. Thus (4) is not universally consistent.

6 Conclusion

As argued in e.g. [2,8,9,10] representing beliefs of agents should be done *syntactically* through predicates of first-order logic to ensure sufficient expressivity. Unfortunately it turns out that representing beliefs syntactically easily leads to inconsistency of the representing system [7,8,13]. This calls for work in finding consistent ways to treat beliefs syntactically, that is, to find restricted ways of representing beliefs in first-order logic such that consistency will necessarily be retained. Some of the most important previous results in this direction can be found in [3,8,9]. In this article we have reached a result which generalizes both [9] and [3]. This result shows that as long as we refrain from expressing quantified negated beliefs (as e.g. in $\forall x\,(B_1(x) \rightarrow \neg B_2(x))$) consistency is always ensured. Our result generalizes [9] by allowing negated beliefs (in [9] negated beliefs like $\neg B_1(\ulcorner \varphi \urcorner)$ will be treated non-classically), and generalizes [3] by allowing quantified beliefs ([3] does not allow quantified beliefs like $\forall x\,(B_1(x) \vee B_2(x))$).

References

[1] Bell, J. L. and M. Machover, "A course in mathematical logic," North-Holland Publishing Co., Amsterdam, 1977, xix+599 pp.

[2] Davis, E., "Representations of Commonsense Knowledge," Morgan Kaufmann, 1990.

[3] des Rivières, J. and H. J. Levesque, *The consistency of syntactical treatments of knowledge*, in: *Theoretical aspects of reasoning about knowledge (Monterey, Calif., 1986)*, Morgan Kaufmann, 1986 pp. 115–130.

[4] Feferman, S., *Toward useful type-free theories I*, The Journal of Symbolic Logic **49** (1984), pp. 75–111.

[5] Gupta, A. and N. Belnap, "The Revision Theory of Truth," MIT Press, 1993.

[6] Lloyd, J. W., "Foundations of logic programming," Springer-Verlag, Berlin, 1987, second edition, xii+212 pp.

[7] Montague, R., *Syntactical treatments of modality, with corollaries on reflection principles and finite axiomatizability*, Acta Philosophica Fennica **16** (1963), pp. 153–166.

[8] Morreau, M. and S. Kraus, *Syntactical treatments of propositional attitudes*, Artificial Intelligence **106** (1998), pp. 161–177.

[9] Perlis, D., *Languages with self-reference I*, Artificial Intelligence **25** (1985), pp. 301–322.

[10] Perlis, D., *Languages with self-reference II*, Artificial Intelligence **34** (1988), pp. 179–212.

[11] Stoltenberg-Hansen, V., I. Lindström and E. R. Griffor, "Mathematical theory of domains," Cambridge University Press, Cambridge, 1994, xii+349 pp.

[12] Tarski, A., *The concept of truth in formalized languages*, in: *Logic, semantics, metamathematics*, Hackett Publishing Co., Indianapolis, IN, 1956 Papers from 1923 to 1938.

[13] Thomason, R. H., *A note on syntactical treatments of modality*, Synthese **44** (1980), pp. 391–395.