

SL: A Subjective, Intensional Logic of Belief¹

Hans Chalupsky and Stuart C. Shapiro

Department of Computer Science
State University of New York at Buffalo
226 Bell Hall
Buffalo, NY 14260
hans@cs.buffalo.edu shapiro@cs.buffalo.edu

Abstract

Logics of belief are usually either quite complex, unintuitive, make overly idealistic assumptions, or all of the above, because they have to cope with the unusual characteristics of the belief operator (relation, predicate). Some of these problematic characteristics are referential opacity, the possible falsehood of objects of belief, belief recursion, identification of referents from outside of the belief operator in quantification contexts, etc. The difficulties faced by traditional logical treatments seem to stem mainly from the fact that an essentially subjective, intensional phenomenon gets analyzed from an objective, outside observer's point of view in an extensional, logical framework.

As an alternative, we propose a subjective, intensional logic **SL**, which takes seriously the usual characterization of belief as a *propositional attitude*, that is, in **SL** belief is treated as a relation between an agent and a proposition (an intensional object). As results we gain technical simplicity and a simple, intuitive semantics for belief sentences.

Introduction

A unifying characteristic of standard logical treatments of knowledge and belief is that they are, in some aspect or other, quite complicated. Syntactic approaches, for example, (Kaplan, 1968; Haas, 1990), usually employ a quotation device which leads to a notationally complex hierarchy of object and metalanguages. Sentential or modal approaches, for example, (Hintikka, 1962; Levesque, 1982; Halpern and Moses, 1992), commonly use a complex and somewhat unintuitive semantic notion, sets of possible worlds, to interpret belief sentences. Referential opacity of belief contexts, quantifying in, possible falsehood of beliefs, or simply technical difficulties with the formalization, make it necessary to complicate things with restricted equality reasoning, standard names, rigid designators, naming maps, etc.

It seems that one of the main sources of these various complexities is that an inherently subjective, intensional

phenomenon such as belief gets analyzed in an objective, extensional way. Notions such as *truth* and *possible worlds* are extensional notions, which are then used to objectively analyze the mental states of believers from an outside observer's point of view. Rectifying assertions of truths about the world with assertions about the mental states of believers that are in some way about this world seems to be the main stumbling block. As an alternative, we propose a subjective, intensional model which we will describe below.

A Logic of Thought

The authors' interest in the subject of belief representation and reasoning results from their long-term goal to build an artificial cognitive agent capable of communicating with other agents in natural language. Such an agent, which from now on we will call *Cassie*, will need what is commonly called a *belief system*, that is some sort of representation and reasoning scheme that describes the set of beliefs held by the agent, and its reasoning with them. Intuitively, Cassie's mind can be thought of as a container that is filled with some sort of "objects", some of which we will call Cassie's beliefs. The question is: What are these objects, and how can they be described?

Following logical tradition, we will pick a formal language whose expressions will denote the objects of Cassie's beliefs. Belief is usually characterized as a *propositional attitude*, that is a relation between an agent and a proposition. Taking this characterization seriously, we will define a language whose expressions denote propositions, and we will construct Cassie's mind as a set of such expressions. Depending on a particular implementation or storage scheme, not all expressions in Cassie's mind will denote propositions actually believed by her. Some might just be residue from pondering certain questions, others might be part of other propositions. To be able to single out Cassie's actual beliefs, we will think of every expression in her mind to be associated with a flag, which, if it is on, will indicate that the proposition denoted by that expression is actually believed by Cassie.

We are of course well aware that in trying to build an artificial cognitive agent we will have to be concerned

¹This is a preliminary version of: Hans Chalupsky and Stuart C. Shapiro, SL: A subjective, intensional logic of belief, in *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pages 165–170, Hillsdale, NJ, August 1994, Lawrence Erlbaum. All quotes should be from, and all citations should be to the published version.

with many other important aspects of agenthood such as plans, actions, intentions, sensory apparatus, etc., however, in this paper we will only talk about the part that deals with beliefs.

It should be evident from the above that we take belief spaces to be sets of propositions rather than sets of sentences. For arguments in favor of this view and against viewing belief as a relation between an agent and a sentence see, for example, (Church, 1950; Shapiro, 1993). By now, the concerned reader might be worried about the fundamental building block of our theory, propositions, since their nature has been subject of much philosophical debate. We will not add to this debate. For us propositions are abstract, intensional entities that are in the domain of discourse, or, as (Creary, 1979, p.176) puts it, they are “abstractions of things psychological.”

Before we go on, two things should be clarified: 1) Our language of propositions will have internal structure and quantification, and 2) even though we model Cassie’s beliefs as a set of expressions denoting the propositions believed by her, our model is not the database approach to belief representation as described and criticized by (Moore, 1977), hence, it does not suffer from the problems of that approach. Figure 1 shows an example snapshot of Cassie’s mind that contains various different kinds of propositions that could be believed by her. Note, that the top-level expressions are Cassie’s immediate beliefs, hence, they need not be framed in an extra application of the belief function B . Only her beliefs about other agent’s beliefs as well as her introspective beliefs have to make use of B . It should be pointed out that despite the standard logical notation all “predicate” names as well as the logical connectives are actually proposition-valued functions. This will become more clear below.

SL: A Fully Intensional, Subjective Logic of Belief

SL is the logic underlying SIMBA. SIMBA, an acronym for simulative belief ascription, is the name of a reasoning system that we are currently developing as part of a forthcoming doctoral dissertation (Chalupsky, 1995). The main emphasis of SIMBA is on the reasoning with other agents’ beliefs, or, put more concretely, Cassie’s reasoning with beliefs she believes are held by other agents.

SL is intended to be the basic building block of the belief system of an artificial cognitive agents such as Cassie. Its language, $L_{\mathbf{SL}}$, is the representational substrate for things in Cassie’s mind, hence, it is a *belief representation language* as opposed to a *world representation language* (see (Maida and Shapiro, 1982, p.296) for a discussion of this distinction). Due to this point of view, the design of $L_{\mathbf{SL}}$ is guided by the following principle assumptions which are derived from (Maida and Shapiro,

1982) and (Shapiro, 1993):

- The domain of discourse, \mathcal{D} , whose elements are denoted by terms of $L_{\mathbf{SL}}$, is a set of *intensional entities* such as propositions, objects of thought, fictional objects, etc.
- \mathcal{D} consists of a set of atomic objects, \mathcal{D}_a , without internal structure, and a set of structured objects, \mathcal{D}_s , whose structure derives in some well-defined way from more primitive objects.
- Two distinct terms of $L_{\mathbf{SL}}$ cannot denote the same entity, a restriction which is called the *uniqueness principle*, because every object is denoted by a unique term.

Since we take intensions or sets of intensional objects as a completely independent realm, as opposed to, for example, the notion used by Montague where intensions are defined in terms of sets of extensions (Montague, 1974), we call our approach *fully intensional* (see (Shapiro and Rapaport, 1987) for more discussion on this view of intensionality). One reason for this approach is that it provides elegant solutions to problems of indirect reference such as McCarthy’s telephone number problem (McCarthy, 1979; Maida and Shapiro, 1982), or problems of representation with *de re* and *de dicto* belief reports (Rapaport, 1986).

Since $L_{\mathbf{SL}}$ is intended to be Cassie’s language of thought in which she represents and reasons, we call **SL** a *subjective* logic. Note, that in Cassie’s representations of other agents’ beliefs, imputations of her own view are a necessary consequence of subjectivity: Cassie can only think in her *own* language of thought, hence, she can only represent and understand other agents’ beliefs in her own terms (this intuition also provides a strong reason against the use of a syntactic approach).

Finally, the uniqueness principle is not so much a requirement than a reflection of how we view cognitive function: Whenever Cassie creates a new (mental) individual (of $L_{\mathbf{SL}}$) for some object of \mathcal{D} , she does so because no already existing individual denotes that object for her, hence, it must be an intensionally different object. Had it been the same, the preexisting individual would have been used as a consequence of cognitive economy.

Formalization of \mathbf{SL}_0

SL is a nonmonotonic logic that supports belief revision, however, at the time of our writing its full formalization is still work in progress; hence, we only present its monotonic precursor which we will call \mathbf{SL}_0 .

The core idea in the formalization of \mathbf{SL}_0 is derived from (Shapiro, 1993): It is that $L_{\mathbf{SL}_0}$ is primarily a language of terms, not of sentences. The main vehicle for constructing these terms are proposition-valued functions. Logical connectives, the main sentence constructors in standard logical treatments, are just a special subset of these functions. Sentences of $L_{\mathbf{SL}_0}$ are only

Cassie

| |
|---|
| !Loves(John, Mary) |
| !∀x (Man(x) ⇒ Mortal(x)) |
| !B(Mary, B(Sally, Loves(John, Sally))) |
| !B(I, Loves(John, Mary)) |
| !¬B(I, Equiv(P, NP)) |
| !B(John, Nice(Mary)) ∨ B(John, ¬Nice(Mary)) |
| !Smart(I) |
| !∀a (B(a, Exists(Santa)) ⇒ Child(a)) |
| P ∧ ¬P |

...plain belief
 ...quantified belief
 ...nested belief
 ...positive introspection
 ...negative introspection
 ...disjunctive belief
 ...false belief
 ...agent quantification
 ...conception without belief

Figure 1: A snapshot of Cassie’s mind: **B** is the name of the belief function and **I** is Cassie’s self concept. The intended interpretations of the other symbols should be evident. The ‘!’ is used to flag the propositions believed by Cassie.

used to associate a particular term with an agent such as Cassie. Consequently, and hopefully not surprisingly, we do not deal with truth values. An expression such as $B(\text{Mary}, B(\text{Sally}, \text{Cute}(\text{John})))$ is then simply a nested function application which yields a proposition as its result, and not a higher-order expression as it would be in a standard treatment where **B** is viewed as a relation.

Notational conventions: We will use **sans serif** for object-language terms, *italics* for meta-variables ranging over such terms, **bold sans serif** for domain objects, and **bold italics** for meta-variables ranging over such objects. The denotation relation is expressed with the usual double brackets, for example, $\llbracket \text{Man}(\text{Hans}) \rrbracket = \text{Man}(\text{Hans})$, $\llbracket p_1 \rrbracket = p_1$. $f(x) \downarrow$ will mean that f is defined for x , $f(x) \uparrow$ that it is not.

Syntax

$L_{\mathbf{SL}_0}$ is an internal language, hence, it is or at least can be different for every agent. However, there is a certain structure that we require for every instance of $L_{\mathbf{SL}_0}$. For that reason, we describe the various sets of symbols defined below as sets of metavariables. Where appropriate, typical object-language instances of these variables are given.

The atoms of $L_{\mathbf{SL}_0}$ are defined as the following sets of symbols:

1. $P =_{def} \{ (,), , , \forall, \exists, ! \}$, punctuation, quantifiers, assertion
2. $I =_{def} \{ i_1, i_2, i_3, \dots \}$, the set of individual constants, e.g., **Mary**, **B₁**, **B₂**, ...
3. i_e , the ego or self concept, e.g., **I**
4. $I_p \subset I$, the set of individual propositional constants
5. $F^n =_{def} \{ f_1^n, f_2^n, f_3^n, \dots \}$, the set of n-ary function constants
6. $F =_{def} \bigcup_{n \geq 1} F^n$, the set of all function constants, e.g., **Loves**, **Knows₁**, ...
7. $F_p \subset F$, the set of propositional function constants
8. $F_l =_{def} \{ f_{\neg}, f_{\wedge}, f_{\vee}, f_{\Rightarrow}, f_{\Leftrightarrow} \}$, the set of logical connective functions, $F_l \subset F_p$, e.g., \neg , \wedge , \vee , \Rightarrow , \Leftrightarrow

9. f_B , the belief function, e.g., **B**
10. $V =_{def} \{ x_1, x_2, x_3, \dots \}$, the set of variables, e.g., $x, y_2, \text{dog}, \dots$

The substitution $e_{x/i}$ is defined as the expression obtained from replacing all free occurrences of x in the expression e by i .

The set of terms, T , is defined by the following inductive rules:

1. Every $i \in I$ is a term. If $i \in I_p$ then i is a propositional term.
2. If t_1, \dots, t_n are terms and $f^n \in F$ then $f^n(t_1, \dots, t_n)$ is a function term. If $f^n \in F_p$ then $f^n(t_1, \dots, t_n)$ is a propositional function term.
3. If x is a variable and t is a propositional function term with an individual i as a subterm such that no occurrence of i is in the scope of a quantifier $\forall x$ or $\exists x$, then $\forall x t_{i/x}$ and $\exists x t_{i/x}$ are propositional terms.

Finally, let T_p be the set of propositional terms. Then $S =_{def} \{ t \mid t \in T_p \}$ is the set of all sentences.

The only role of sentences is to flag the set of propositional terms that are actually believed by the agent under consideration. Functions are intended to generate structured names that denote structured individuals, for example, the function symbol **Loves** is intended to denote the function which has all propositions of the form that *one individual loves another individual* as its range. The term **Loves(John, Mary)** is intended to denote the proposition that *John loves Mary*. The term **B(Lucy, Loves(John, Mary))** is intended to denote the proposition that *Lucy believes that John loves Mary*, a proposition which contains the proposition from the previous example as a part.

For now, we assume that the structure defined by the standard syntax of function application and composition is sufficient to describe the structure of objects such as propositions. The main drawback of this scheme is that it always encodes the order of the arguments even if the resulting proposition should be order independent. The

logical-connective functions provide a good example: In our current scheme the terms $\wedge(\mathbf{P}, \mathbf{Q})$ and $\wedge(\mathbf{Q}, \mathbf{P})$ ² denote two different propositions (because of the uniqueness principle), but one could make an argument that this should not be so. There are solutions to this problem, but we will not discuss them here.

Agent Interpretations

Intuitively, Cassie’s belief system is filled with a set of $L_{\mathbf{SL}_0}$ expressions; hence, $L_{\mathbf{SL}_0}$ can be viewed as Cassie’s language of thought. It is an *internal* language, much like the language used inside the scope of Konolige’s modal belief operator $[S_i]$ (Konolige, 1986), but \mathbf{SL}_0 does not define an *external* language that describes Cassie’s belief system from an outside observer’s point of view. Since $L_{\mathbf{SL}_0}$ is Cassie’s language of thought, she uses that very language to represent other agents’ beliefs, and there is no need for quotation, standard names, naming maps, etc.

As builders of Cassie, we are of course interested in how her internal language is linked to the outside. That these internal expressions denote the proper individuals and propositions is a prerequisite for her being able to understand and be understood by other agents. The connection between Cassie’s internal language and the external, though not extensional, domain \mathcal{D} will be made via *agent interpretations*.

Definition: An *intensional domain of discourse*, \mathcal{D} , is a non-empty set of intensional individuals which consists of two disjoint subsets: \mathcal{D}_a , a set of atomic elements, and \mathcal{D}_s , a set of structured elements. Each of these parts is further split into a propositional part, \mathcal{D}_{a_p} and \mathcal{D}_{s_p} which taken together form \mathcal{D}_p , and a non-propositional part, $\mathcal{D}_{a_{np}}$ and $\mathcal{D}_{s_{np}}$ which taken together form \mathcal{D}_{np} .

An intensional domain is intended to contain objects of thought, discourse entities, concepts, propositions, impossible objects, fictional objects, etc., all of which are denoted by terms of $L_{\mathbf{SL}_0}$. For the purposes of our exposition, it suffices to mainly distinguish between propositions and other objects. To provide proper denotations for function terms we need actual functions that operate on \mathcal{D} . Because of the uniqueness principle, not just any set of functions will do. The following definition specifies some necessary characteristics for a set of domain functions that can be used in an interpretation:

Definition: Let \mathcal{D} be an intensional domain of discourse. A set of functions \mathcal{F} is a *basis* for \mathcal{D}_s iff

1. every $\mathbf{f} \in \mathcal{F}$ is an n-ary, injective (or 1-to-1) function $\mathcal{D}^n \rightarrow \mathcal{D}_s$, $n \geq 1$, and
2. for every $\mathbf{f}_1, \mathbf{f}_2 \in \mathcal{F}$ such that $\mathbf{f}_1 \neq \mathbf{f}_2$ the range of \mathbf{f}_1 is disjoint from the range of \mathbf{f}_2 , and

²Usually, the logical connective functions are written in the standard prefix and infix notation (see Figure 1) instead of as function applications.

3. for every $i \in \mathcal{D}_s$ there is an n-ary function $\mathbf{f}^n \in \mathcal{F}$ and a set of arguments $\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{x}_i \in \mathcal{D}$ such that $\mathbf{f}^n(\mathbf{x}_1, \dots, \mathbf{x}_n) = i$.

The conditions that the functions need to be 1-to-1, and that their ranges are required to be mutually disjoint are direct consequences of the uniqueness principle. If a particular function were not 1-to-1, then there would be two different sets of arguments for which the function would have the same value. If the ranges were not disjoint, then there would be two different functions with a value in common. If such a set of functions were to be used in an interpretation then there would be at least two different terms of $L_{\mathbf{SL}_0}$ with the same denotation, a violation of the uniqueness principle.

Finding an actual basis of domain functions is a very hard and mainly unsolved problem. It amounts to the construction of a theory about how natural language gets mapped into propositional meaning representations. Every domain function can be viewed as a kind of case frame, and the problem then becomes to find a correct and complete set of such case frames. This problem is not unique to our approach, every standard logical approach faces a similar problem as soon as an actual domain theory has to be constructed.

The denotations of the logical connective functions and the belief function are the only ones defined by \mathbf{SL}_0 itself, as opposed to leaving that up to a particular interpretation:

Definition: The set of *logical connective functions*, \mathcal{F}_l , is the set $\{\mathbf{f}_{\neg}, \mathbf{f}_{\wedge}, \mathbf{f}_{\vee}, \mathbf{f}_{\Rightarrow}, \mathbf{f}_{\Leftrightarrow}\}$ whose individual elements are defined as follows (let $\mathbf{p}, \mathbf{q} \in \mathcal{D}_p$ be arbitrary propositions):

- $\mathbf{f}_{\neg} : \mathcal{D}_p \rightarrow \mathcal{D}_{s_p}$. The value of $\mathbf{f}_{\neg}(\mathbf{p})$ is the proposition that it is not the case that \mathbf{p} .
- $\mathbf{f}_{\wedge} : \mathcal{D}_p^2 \rightarrow \mathcal{D}_{s_p}$. The value of $\mathbf{f}_{\wedge}(\mathbf{p}, \mathbf{q})$ is the proposition that it is the case that \mathbf{p} and \mathbf{q} .
- $\mathbf{f}_{\vee} : \mathcal{D}_p^2 \rightarrow \mathcal{D}_{s_p}$. The value of $\mathbf{f}_{\vee}(\mathbf{p}, \mathbf{q})$ is the proposition that it is the case that either \mathbf{p} or \mathbf{q} or both.
- $\mathbf{f}_{\Rightarrow} : \mathcal{D}_p^2 \rightarrow \mathcal{D}_{s_p}$. The value of $\mathbf{f}_{\Rightarrow}(\mathbf{p}, \mathbf{q})$ is the proposition that if it is the case that \mathbf{p} then it is the case that \mathbf{q} .
- $\mathbf{f}_{\Leftrightarrow} : \mathcal{D}_p^2 \rightarrow \mathcal{D}_{s_p}$. The value of $\mathbf{f}_{\Leftrightarrow}(\mathbf{p}, \mathbf{q})$ is the proposition that it is the case that \mathbf{p} if and only if it is the case that \mathbf{q} .

Definition: The *belief function*, $\mathbf{f}_{\mathbf{B}}$, is defined as follows (let $\mathbf{a} \in \mathcal{D}_{a_{np}}, \mathbf{p} \in \mathcal{D}_p$):

- $\mathbf{f}_{\mathbf{B}} : \mathcal{D}_{a_{np}} \times \mathcal{D}_p \rightarrow \mathcal{D}_{s_p}$. The value of $\mathbf{f}_{\mathbf{B}}(\mathbf{a}, \mathbf{p})$ is the proposition that \mathbf{a} believes \mathbf{p} .

Since we do not have a formal theory of propositions (our semantic domain), we used English sentences to define the class of propositions yielded by the proposition-valued functions defined above. So, in a sense, English is our external language that we use to interpret Cassie’s internal language. This is not all that much different

from standard truth recursion rules which usually define the truth of compound expressions with help of natural-language conjunctions whose semantic function is taken to be understood.

Terms of $L_{\mathbf{SL}_0}$ will be mapped onto the domain of discourse by means of an interpretation function which has to fulfill the following requirements:

Definition: Let \mathcal{D} be an intensional domain of discourse and \mathcal{F} a basis for \mathcal{D}_s . A function $int : T \rightarrow \mathcal{D}$ is an *admissible $L_{\mathbf{SL}_0}$ interpretation function* for \mathcal{D} and \mathcal{F} if it satisfies the following conditions:

1. If $i \in I$ then $int(i) \in \mathcal{D}_a$.
2. If $i \in I_p$ then $int(i) \in \mathcal{D}_{a_p}$.
3. If $f \in F$ then $int(f) \in \mathcal{F}$.
4. If $f \in F_p$ then $int(f) \in \mathcal{F}_p$.
5. $int(\neg) = \mathbf{f}_{\neg}$, $int(\wedge) = \mathbf{f}_{\wedge}$, $int(\vee) = \mathbf{f}_{\vee}$,
 $int(\Rightarrow) = \mathbf{f}_{\Rightarrow}$, $int(\Leftrightarrow) = \mathbf{f}_{\Leftrightarrow}$, $int(\mathbf{B}) = \mathbf{f}_{\mathbf{B}}$.
6. If t is an n-ary function term of the form $f^n(x_1, \dots, x_n)$ then $int(t) = int(f^n)(int(x_1), \dots, int(x_n))$.
7. If t is a term of the form $\forall x t'$ then $int(t)$ is the proposition that for every atomic domain element $int(i)$ such that $int(t'_{x/i}) \downarrow$ it is the case that $int(t'_{x/i})$.
8. If t is a term of the form $\exists x t'$ then $int(t)$ is the proposition that there is an atomic domain element $int(i)$ such that $int(t'_{x/i}) \downarrow$ and it is the case that $int(t'_{x/i})$.
9. For any two terms t_1, t_2 if $t_1 \neq t_2$ then $int(t_1) \neq int(t_2)$.

Now we can finally define how $L_{\mathbf{SL}_0}$ terms in the “mind” of an agent are to be interpreted:

Definition: Let $a = \langle \mathcal{D}, \mathcal{F}, \mathcal{B}, e, int \rangle$ be an *agent interpretation structure*, where \mathcal{D} is an intensional domain of discourse, \mathcal{F} is a basis for \mathcal{D}_s , $\mathcal{B} \subset \mathcal{D}_p$ is the base set of the agent’s beliefs, $e \in \mathcal{D}_{a_{n_p}}$ is the agent’s ego or self-concept, and int is an admissible $L_{\mathbf{SL}_0}$ interpretation function for \mathcal{D} and \mathcal{F} . Then for every $t \in T$: $\llbracket t \rrbracket_a =_{def} int(t)$.

The agent’s base set of beliefs \mathcal{B} is intended to describe the set of propositions believed by the agent without being justified by the logic \mathbf{SL}_0 itself, i.e., they serve as a set of extra-logical belief axioms. Intuitively, such beliefs are formed from sensory information, from being told by somebody, etc. Since we are concerned with the modeling of realistic agents, we will usually take \mathcal{B} to be finite.

Apart from being able to interpret what actual propositions are believed by an agent, it would be interesting to know whether the agent behaves rationally. One aspect of rational behavior is to draw proper conclusions from one’s current set of beliefs. Below we present a notion of *justification* which defines whether an agent is justified in believing a certain proposition relative to some agent interpretation. We will use the notation $\mathbf{J}_{\bar{a}}!p$ (a turnstile with a J-bar for ‘justification’) to express that an agent is justified in believing $\llbracket p \rrbracket_a$ relative to some

interpretation a . Intuitively, if Cassie tells us that she believes some proposition $\llbracket p \rrbracket$, then we can verify whether she is justified in believing it by checking whether $\mathbf{J}_{\bar{a}}!p$ holds. If Cassie tells us that she does not believe $\llbracket p \rrbracket$, then checking whether $\mathbf{J}_{\bar{a}}!p$ is not of much help, because we do not assume that Cassie’s beliefs are deductively closed. In that sense, our notion of justification characterizes the reasoning of an ideal agent, or, to use Levesque’s terminology, it specifies the set of *implicit* beliefs of Cassie (Levesque, 1984).

Definition: Let t, t_1, t_2 be propositional terms of $L_{\mathbf{SL}_0}$, $a = \langle \mathcal{D}, \mathcal{F}, \mathcal{B}, e, int \rangle$ be an agent interpretation structure, and i_e be the $i \in I$ such that $\llbracket i \rrbracket_a = e$. Then we define whether an agent is justified in believing $\llbracket t \rrbracket_a$ relative to an agent model, written $\mathbf{J}_{\bar{a}}!t$. To avoid a circular definition we do this in two steps. First we define $\mathbf{J}_{\bar{a}}^{\circ}!t$ which does not say anything about belief sentences that involve other agents:

1. If $\llbracket t \rrbracket_a \in \mathcal{B}$ then $\mathbf{J}_{\bar{a}}^{\circ}!t$.
2. $\mathbf{J}_{\bar{a}}^{\circ}!\neg t$ iff $\mathbf{J}_{\bar{a}}^{\circ}!t$.
3. $\mathbf{J}_{\bar{a}}^{\circ}!t_1 \wedge t_2$ iff $\mathbf{J}_{\bar{a}}^{\circ}!t_1$ and $\mathbf{J}_{\bar{a}}^{\circ}!t_2$.
4. $\mathbf{J}_{\bar{a}}^{\circ}!t_1 \vee t_2$ iff $\mathbf{J}_{\bar{a}}^{\circ}!t_1$ or $\mathbf{J}_{\bar{a}}^{\circ}!t_2$.
5. $\mathbf{J}_{\bar{a}}^{\circ}!t_1 \Rightarrow t_2$ iff $\mathbf{J}_{\bar{a}}^{\circ}!t_1$ or $\mathbf{J}_{\bar{a}}^{\circ}!t_2$.
6. $\mathbf{J}_{\bar{a}}^{\circ}!\forall x t$ iff $\mathbf{J}_{\bar{a}}^{\circ}!t_{x/i}$ for all $i \in I$ such that $\llbracket t_{x/i} \rrbracket_a \downarrow$.
7. $\mathbf{J}_{\bar{a}}^{\circ}!\exists x t$ iff $\mathbf{J}_{\bar{a}}^{\circ}!t_{x/i}$ for some $i \in I$ such that $\llbracket t_{x/i} \rrbracket_a \downarrow$.
8. $\mathbf{J}_{\bar{a}}^{\circ}!\mathbf{B}(i_e, t)$ iff $\mathbf{J}_{\bar{a}}^{\circ}!t$.

Now we can define $\mathbf{J}_{\bar{a}}!t$:

9. $\mathbf{J}_{\bar{a}}!\mathbf{B}(b, t)$, $b \neq i_e$ iff there are terms p_1, \dots, p_n , $n \geq 0$ such that $\mathbf{J}_{\bar{a}}!\mathbf{B}(b, p_1), \dots, \mathbf{J}_{\bar{a}}!\mathbf{B}(b, p_n)$ and $\mathbf{J}_{\bar{a}'}!t$ with $a' = \langle \mathcal{D}, \mathcal{F}, \{\llbracket p_i \rrbracket_a, \dots, \llbracket p_n \rrbracket_a\}, \llbracket b \rrbracket_a, int \rangle$.
10. Otherwise: $\mathbf{J}_{\bar{a}}!t$ iff $\mathbf{J}_{\bar{a}}^{\circ}!t$.

The cases for the logical connectives in the definition above are very similar to the truth recursion rules used in standard semantics of first-order predicate logic. The main difference is the terminology, because we are not concerned with the notion of truth. The interesting cases are discussed below:

Case 1 deals with the fact that our interpretation function simply maps propositional terms onto propositions. It does not say anything about the agent’s belief status regarding these propositions. By comparison, the corresponding case in a standard first-order semantics is the one that handles the truth of ground sentences such as $\mathbf{Red}(\mathbf{Apple}_1)$. There the interpretation function maps predicate or relation names onto sets, and the truth of atomic sentences is determined by membership of the interpreted arguments in these sets.

Case 8 defines the semantics of introspection. It exemplifies best the subjective character of \mathbf{SL}_0 , since the self belief is based on a “plain” proposition that is not itself nested inside a belief function (though it could be). This case has some similarity with a subjective version of the inference rule S4 of modal systems, but of course,

S4 is an inference rule of the *deductive system* of certain modal logics, while the above is a definition of the *semantics* of introspection in \mathbf{SL}_0 . We do not need an extra case $\mathbb{J}_a^\circ !\neg B(i_e, t)$ iff $\mathbb{J}_a^\circ !t$ to characterize negative introspection, because that follows as a simple theorem from the definition.

Case 9 defines the semantics of a mechanism called *simulative reasoning* (Creary, 1979; Chalupsky, 1993), in which an agent such as Cassie hypothetically assumes beliefs it thinks are held by other agents, and then tries to infer consequences from these beliefs with its own reasoning mechanism. The result of such a simulation gets then ascribed to the simulated agent. The semantics of this is captured by basing the justification of a belief sentence on a variant of the agent model which only contains the beliefs of the simulated agent as the base set, and which uses the simulated agent as its ego.

There are usually infinitely many interpretations under which a particular sentence is justified, hence, we normally work with the following stronger definition of justification:

Definition: $!p_1, \dots, !p_n \mathbb{J} !q, n \geq 0$ iff all agent interpretations which justify $!p_1, \dots, !p_n$ also justify $!q$.

Since \mathcal{B} might contain arbitrary propositions, it is also useful to only consider *consistent* agent interpretations. An interpretation a is consistent if there is no p such that $\mathbb{J}_a !p \wedge \neg p$. An alternative to the consistency requirement would be to restrict \mathcal{B} to primitive propositions which do not contain any logical connective functions or quantifiers, however, this would not allow us to model agents who believe in universally quantified propositions that they were simply told about.

Discussion

Did we achieve our goal? Is \mathbf{SL}_0 less complicated than standard logics of belief but at least as expressive, and does it have a more intuitive semantics of belief sentences? People have of course different views on what counts as simple or intuitive, but let us quickly discuss some of these points:

Suppose that Cassie believes that John loves Mary, and that Sally believes that John loves Mary. In a first-order syntactic logic the above would be expressed with help of some sort of quotation device, e.g., $\text{Loves}(\text{John}, \text{Mary})$ and $B(\text{Sally}, \text{!Loves}(\text{John}, \text{Mary})^1)$. This becomes more complicated with deeper nesting and once quantification into quotation contexts is considered. It also complicates the proof theory. More serious than the syntactic complications are the semantic implications. Both sentences denote truth values. The first sentence describes a relation between two agents, the second describes a relation between an agent and a sentence, i.e., a syntactic object which basically is a complicated string constant. For an agent that uses this as its language of thought the expressions $\text{Loves}(\text{John}, \text{Mary})$ and $\text{!Loves}(\text{John}, \text{Mary})^1$ are of very different character, they

mean completely different things. In (Shapiro, 1993) it is argued that to the agent $B(\text{Sally}, \text{!Loves}(\text{John}, \text{Mary})^1)$ actually means that Sally believes some incomprehensible gibberish, because the object of the belief is expressed in a language different from its own language of thought.

Modal logics of belief usually employ a different modal operator for every agent. These operators get applied to complete sentences, e.g., $B_{\text{Sally}}\text{Loves}(\text{John}, \text{Mary})$. There is no need for quotation, but we have the disadvantage that we cannot express sentences that quantify over agents. Moreover, the semantics of such modal belief sentences is usually rendered as something like the following: $\text{Loves}(\text{John}, \text{Mary})$ holds in all worlds that Sally considers possible according to some accessibility relation. While such Kripke-style possible-worlds semantics are technically elegant, they are certainly not among the most intuitive explanations of the concept of belief. Another drawback of standard possible-worlds analyses is that agents are modeled as logically omniscient.

In \mathbf{SL}_0 the above situation would be represented with $!\text{Loves}(\text{John}, \text{Mary})$ and $!B(\text{Sally}, \text{Loves}(\text{John}, \text{Mary}))$. The semantics of the propositional term $\text{Loves}(\text{John}, \text{Mary})$, or $\llbracket \text{Loves}(\text{John}, \text{Mary}) \rrbracket$, is the proposition that *John loves Mary*. Once one accepts that it is as easy to accept that $\llbracket B(\text{Sally}, \text{Loves}(\text{John}, \text{Mary})) \rrbracket$ is the proposition that *Sally believes that John loves Mary*, a proposition that contains the previous one as its part; hence, the semantics of belief sentences is exactly the same and as simple as the semantics of plain sentences (which are Cassie's immediate beliefs). This model also meshes well with the characterization of belief as a propositional attitude. Since \mathbf{SL}_0 is subjective, all constants are in Cassie's head, or put differently, the only way Cassie can think about other agents' beliefs is in her own terms or language of thought. Therefore there is no need for technical devices such as standard names, rigid designators, naming maps, etc. in order to identify objects across opaque belief spaces of agents, or across possible worlds.

Because of space restrictions, we did not present the inference mechanism (or proof theory) of \mathbf{SL}_0 . The current implementation of SIMBA which is built upon \mathbf{SL}_0 uses a natural-deduction-based inference package to perform reasoning. Cassie's reasoning with the beliefs of other agents is based on a simulative reasoning mechanism.

\mathbf{SL}_0 does of course have shortcomings which will be overcome by its full-blown, nonmonotonic version \mathbf{SL} . One of them is a form of *simulative idealization*, where Cassie assumes every agent to make the same inferences as she does in its simulation, hence, she cannot deal with a situation where she believes $B(\text{Mary}, P)$, $B(\text{Mary}, P \Rightarrow Q)$, and $\neg B(\text{Mary}, Q)$. Another shortcoming is that \mathbf{SL}_0 cannot handle inconsistent agents, for example, if Cassie believes $B(\text{Mary}, P)$, $B(\text{Mary}, P \Rightarrow Q)$

and $B(\text{Mary}, P \Rightarrow \neg Q)$. Such situations can occur if Cassie's model of another agent is incorrect, or if that agent really is inconsistent. Inconsistencies of this kind need to be handled gracefully without jeopardizing Cassie's own reasoning.

References

- Chalupsky, H. (1993). Using hypothetical reasoning as a method for belief ascription. *Journal of Experimental and Theoretical Artificial Intelligence (JETAI)*, 5(2&3):119–133.
- Chalupsky, H. (1995). Belief ascription by way of simulative reasoning. forthcoming PhD dissertation.
- Church, A. (1950). On Carnap's analysis of statements of assertion and belief. *Analysis*, 10:97–99.
- Creary, L. G. (1979). Propositional attitudes: Fregean representations and simulative reasoning. In *Proceedings of the Sixth International Conference on Artificial Intelligence*, pages 176–181, Palo Alto, CA. Morgan Kaufmann.
- Haas, A. R. (1990). Sentential semantics for propositional attitudes. *Computational Linguistics*, 16:213–233.
- Halpern, J. Y. and Moses, Y. O. (1992). A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54(3):319–379.
- Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca, NY.
- Kaplan, D. (1968). Quantifying in. *Synthese*, 19:178–214.
- Konolige, K. (1986). *A Deduction Model of Belief*. Morgan Kaufmann, Palo Alto, CA.
- Levesque, H. J. (1982). A formal treatment of incomplete knowledge bases. Technical Report 614, Fairchild. FLAIR Technical Report No. 3.
- Levesque, H. J. (1984). A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 198–202, Palo Alto, CA. Morgan Kaufmann.
- Maida, A. S. and Shapiro, S. C. (1982). Intensional concepts in propositional semantic networks. *Cognitive Science*, 6(4):291–330. Reprinted in R. J. Brachman and H. J. Levesque, eds. *Readings in Knowledge Representation*, Morgan Kaufmann, Los Altos, CA, 1985, 170–189.
- McCarthy, J. (1979). First order theories of individual concepts and propositions. In Meltzer, B. and Michie, D., editors, *Machine Intelligence 9*, pages 120–147. Ellis Horwood.
- Montague, R. (1974). Universal grammar. In Thomason, R. H., editor, *Formal Philosophy: Selected Papers of Richard Montague*, chapter 7, pages 222–246. Yale University Press.
- Moore, R. C. (1977). Reasoning about knowledge and action. In *Proceedings of the Fifth International Conference on Artificial Intelligence*, pages 223–227, Palo Alto, CA. Morgan Kaufmann.
- Rapaport, W. J. (1986). Logical foundations for belief representation. *Cognitive Science*, 10:371–422.
- Shapiro, S. C. (1993). Belief spaces as sets of propositions. *Journal of Experimental and Theoretical Artificial Intelligence (JETAI)*, 5(2&3):225–235.
- Shapiro, S. C. and Rapaport, W. J. (1987). SNePS considered as a fully intensional propositional semantic network. In Cercone, N. and McCalla, G., editors, *The Knowledge Frontier*, pages 263–315. Springer-Verlag, New York.