

Belief Spaces

Wouter Beek & Remko Scha

December 3, 2008

1 Opacity

The language of first-order logic, that we use to interpret natural language sentences, is an **extensional language**. In an extensional language the meaning of an expression is the thing that the expression denotes. Therefore, if two expressions have the same extension, i.e. if they denote the same thing, then they can be freely intersubstituted. The intersubstitutability of expressions having the same extension/denotation is called the *extensionality principle*.

Theorem 1.1 (Extensionality Principle). $(\chi = \chi') \models (\phi \leftrightarrow \phi[\chi/\chi'])$

According to the extensionality principle we can replace an expression χ with any other expression χ' having the same denotation. We can perform this intersubstitution in any sentence ϕ containing one or more occurrences of χ . Assume, for instance, that the following sentence is true.

Biggles is the thief. (1)

In this case, we can replace ‘Biggles’ for ‘the thief’ in the following sentence:

The thief entered through the window. (2)

This results in:

Biggles entered through the window. (3)

In other words, sentences (2) and (3) have the same truth-values in all models that satisfy sentence (1).

However, the principle of extensionality does not hold for all natural language sentences. More specifically, it does not hold for sentences containing the verb ‘believe’. For even though (1) holds, the following two sentences need not have the same truth-values:

The detective believes that the thief entered through the window. (4)

The detective believes that Biggles entered through the window. (5)

It can be the case that the detective believes that the thief entered through the window, but does not know that Biggles entered through the window. The detective might not even know a man named ‘Biggles’ at all, or he might know Biggles but be unaware that Biggles is the thief. The reason for the breaking down of the principle of extensionality is clear: the detective does not have total knowledge of the world. Most notably, the sentence “Biggles is the thief” might be missing from the detective’s knowledge.

2 Defining semantics and pragmatics for belief

Let us take the following example sentence:

John believes that Mary walks (6)

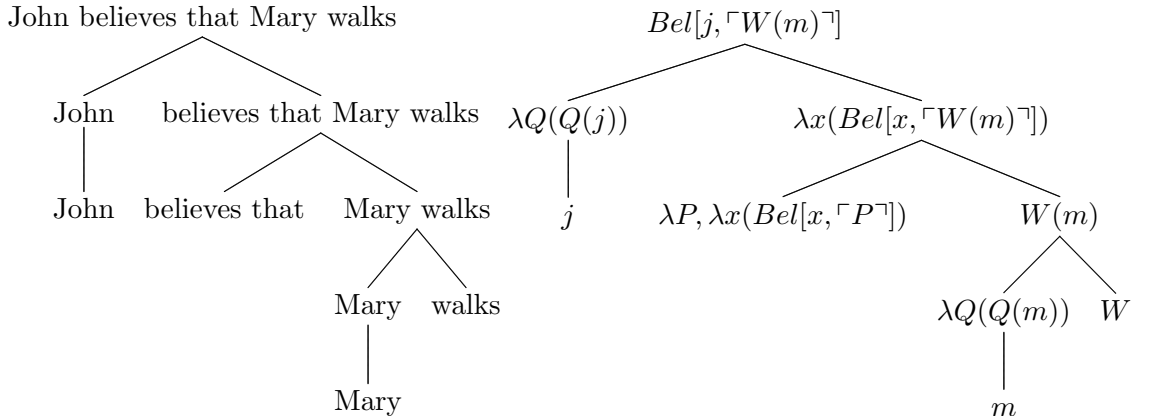
We add a special semantic rule for the verb ‘believe’, in which we treat of the expression ‘believes that’ as a syncategorematic term that is concatenated to a VP . This means that it is introduced as a term that does not have an individual meaning. It only has meaning when occurring in combination with a VP .

Semantics 2.1 (Belief).

$[[\text{believes that}, VP]] = \lambda x(Bel[x, \ulcorner VP \urcorner])$, and is of type $\langle t, \langle e, t \rangle \rangle$.

The difference between ϕ and $\ulcorner \phi \urcorner$ is that while in the former case we can intersubstitute ϕ with any formula ψ that has the same truth value as ϕ , in the latter case this is not allowed.

The sample sentence 6 can then be analyzed as follows:



We will represent a person a 's beliefs as a model satisfying those beliefs. E.g. if a person only believes that every man loves a woman and that john is a man and mary a woman¹, then this person's belief is represented as $\langle \{d_1, d_2\}, \{ \langle john, d_1 \rangle, \langle mary, d_2 \rangle, \langle man, \{d_1\} \rangle, \langle woman, \{d_2\} \rangle, \langle love, \{ \langle d_1, d_2 \rangle \} \} \rangle$.

Since no person can be considered to have complete knowledge of the world, this model must necessarily be partial. This means that there are many different models representing the beliefs of one person. The partial model we choose is the minimal partial model satisfying the things believed by a . The minimal partial domain is the domain containing only entities that are necessary for satisfying a 's beliefs. This is why the domain is the above example is $\{d_1, d_2\}$ and not e.g. $\{d_1, d_2, d_3\}$.

The minimal partial interpretation function only contains interpretations for predicates and constants that occur in a 's beliefs, and (ii) the extensions for each of the predicates is naturally constrained by the definition of the minimal domain, i.e. only tuples are allowed that contain entities belonging to the partial domain exclusively. This is why in the above example we do not have $\langle human, \{d_1, d_2\} \rangle$, nor $\langle love, \{ \langle d_1, d_2 \rangle, \langle d_1, d_3 \rangle \} \rangle$ as part of the interpretation function.

It is clear that $Bel[a, \ulcorner \phi \urcorner]$ cannot be a formula in first-order logic, since Bel is not a relation between entities and truth-values. We define $f_{\mathcal{M}} : D \rightarrow \{ \mathcal{M}' \mid \mathcal{M}' \text{ is a partial model} \}$. This function $f_{\mathcal{M}}$ thus maps the entities from the domain onto models. We shall further add the restriction that the function assigns a model to those entities in the domain that are rational agents, and returns the tuple $\langle \emptyset, \emptyset \rangle$ for all the other entities.²

We shall now interpret the semantic operator Bel .

Semantics 2.2 (Belief).

For any formula of the form $Bel[a, \ulcorner \phi \urcorner]$ we have:

$$\| Bel[a, \ulcorner \phi \urcorner] \|_{\mathcal{M}} = 1 \text{ iff } f_{\mathcal{M}}(a) \models \phi$$

In the formula $Bel[a, \ulcorner \phi \urcorner]$, the logical expression a is of type e . The logical expression ϕ is of type t . The difference between ϕ and $\ulcorner \phi \urcorner$ is that while in the former case we can intersubstitute ϕ with any formula ψ that has the same truth value as ϕ , in the latter case this is not allowed.

For the sample sentence (6) we then get:

¹This is of course a very small set of beliefs, and it is very unlikely – if not impossible – that only these hold for a person. But the purpose here is to illustrate how beliefs are represented, and not to give an empirically adequate account of a real person's belief space.

²It will become clear that, as a consequence of defining $f_{\mathcal{M}}$ thus, that all entities (including the non-rational ones) believe any tautology.

$$\begin{aligned}
& \|Bel[j, \ulcorner W(m) \urcorner]\| = 1 & (7) \\
& \text{iff } f(j) \models W(m) \\
& \text{iff } \forall \phi \in f(j). \|\phi\| = 1 \Rightarrow \|W(m)\| = 1
\end{aligned}$$

3 The nesting of belief

Up to now we have only seen belief-operators at a nesting of one level deep, e.g. as in “John believes that Mary walks.” But there are also sentences that contain nestings of belief, like the following:

$$\text{Peter believes that John believes that Mary walks.} \quad (8)$$

In this case, it does not suffice to know what the belief spaces of both Peter and John are. What is more, the meaning of this sentence cannot involve reference to the belief space of John at all, for it is possible for Peter to believe that John believes something that he, in fact, does not believe. We therefore need the iterated model of Peter’s belief about John’s belief. So instead of \mathcal{M}_j (John’s belief) and \mathcal{M}_p (Peter’s belief), we need the inductive model $\mathcal{M}_{p,j}$, representing Peter’s belief about John’s belief. This notion of an iterated model can be defined inductively:

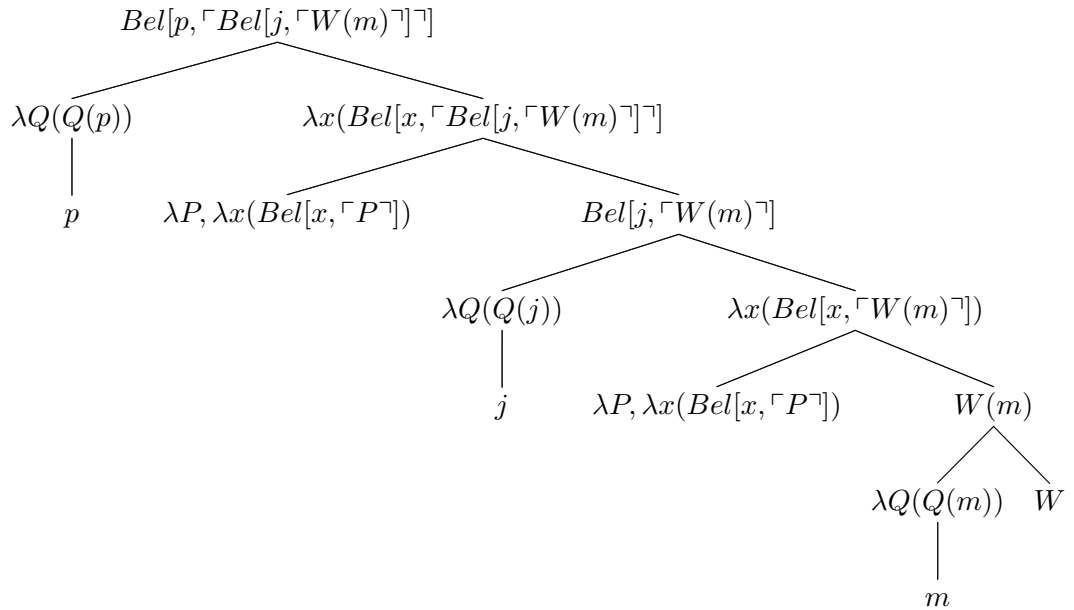
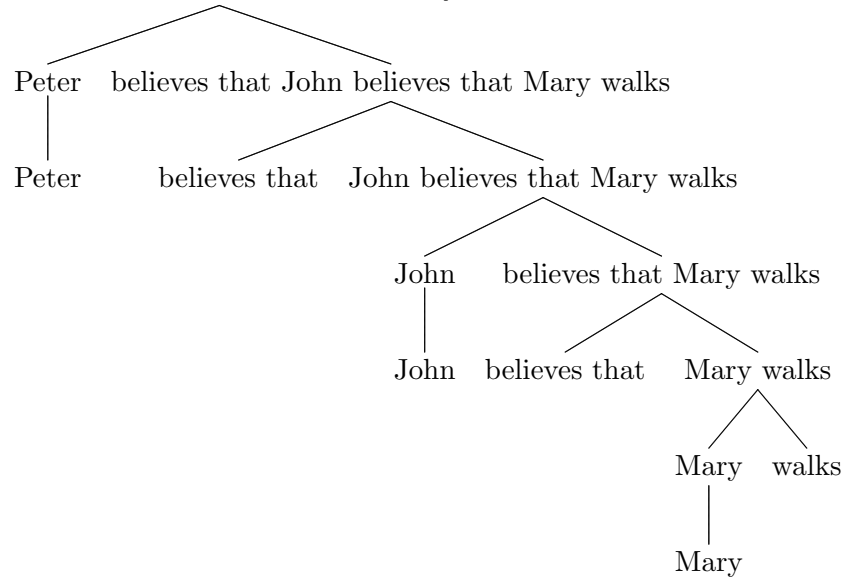
Definition 3.1 (Iterated models).

- If \mathcal{M}_{a_i} is a model, then it is an iterated model.
- If $\mathcal{M}_{a_1, \dots, a_n}$ is an iterated model, and \mathcal{M}_b is a model, then $\mathcal{M}_{a_1, \dots, a_n, b}$ is an iterated model.

Function $f_{\mathcal{M}}$ is extended in order to map tuples of agents onto iterated models. If a non-rational entity occurs somewhere in the tuple, the function gives $\langle \emptyset, \emptyset \rangle$.

It is important to note that an iterated model is the same kind of thing as a normal model. Moreover, the order in the agents in the subscript does not matter. E.g. what John believes about Peter (i.e. $\mathcal{M}_{j,p}$) need not be the same as what Peter believes about John (i.e. $\mathcal{M}_{p,j}$). If we use the semantic rule for belief 2.1 on sample sentence 6, we get the following:

Peter believes that John believes that Mary walks



We must alter the pragmatic rule 2.2 a bit in order to make it apply to the case of nested beliefs too.

Semantics 3.1 (Iterated belief).

For any ‘formula’ of the form $Bel[a_1, Bel[a_2, \dots Bel[a_n, \phi] \dots]]$ we have that:

$\|Bel[a_1, \lceil Bel[a_2, \dots Bel[a_n, \lceil \phi \rceil] \dots \rceil] \rceil\| = 1$ iff $f_{\mathcal{M}}(\langle a_1, \dots, a_n \rangle) \models \phi$.

In the formula $Bel[a_1, \lceil Bel[a_2, \dots Bel[a_n, \lceil \phi \rceil] \dots \rceil]$, the logical expressions a_i are of type e ; the logical expression ϕ is of type t ; and $f_{\mathcal{M}}(\langle a_1, \dots, a_n \rangle) = \mathcal{M}_{a_1, \dots, a_n}$.

We can now analyze sentence (8) as:

$$\begin{aligned} \|Bel[p, \lceil Bel[j, \lceil W(m) \rceil] \rceil] \| &= 1 & (9) \\ \text{iff } f_{\mathcal{M}}(\langle p, j \rangle) &\models W(m) \\ \text{iff } \forall \phi \in \mathcal{M}_{p, j}. \|\phi\| = 1 &\Rightarrow \|W(m)\| = 1 \end{aligned}$$

There is still a problem here, since the nestings of belief operators need not occur in such a nice and orderly fashion as was the case in sentence (8). What about sentences like ‘‘John believes that Mary doesn’t walk and that Peter believes that Mary does walk’’, analyzed as $Bel[j, \lceil \neg W(m) \wedge Bel[p, \lceil W(m) \rceil] \rceil]$? We must either alter the semantic rule for ‘believe’ so that it accounts for these alternative nestings too; or we leave the semantic rule as it is and introduce a canonization procedure that gives, for every belief-sentence, the corresponding canonical form to which the semantic rule can be applied. We choose for the latter option (although the former might with equally good reasons have been attempted).

Definition 3.2 (Canonical form).

- $B[a, \lceil \neg \phi \rceil]$ translates into $\neg B[a, \lceil \phi \rceil]$.
- $B[a, \lceil \phi \wedge \psi \rceil]$ translates into $B[a, \lceil \phi \rceil]$ and $B[a, \lceil \psi \rceil]$.³
- If none of the above conversion rules apply, then the belief-expression has a canonical form.⁴

³The other connectives are left out here, their translations can be easily derived by $\phi \vee \psi := \neg(\neg\phi \wedge \neg\psi)$, $\phi \rightarrow \psi := \neg\phi \vee \psi$, and $\phi \leftrightarrow \psi := (\phi \rightarrow \psi) \wedge (\psi \rightarrow \phi)$.

⁴We have not included the canonization rules for quantifiers here. The reason for this is that the semantics for arbitrary occurrences of quantifiers are a bit tricky. Observe the following sentence:

John believes that there is something Mary believes to be walking. (10)

$$\text{Bel}[j, \lceil \exists x. Bel[m, \lceil W(x) \rceil] \rceil] \quad (11)$$

This is not the same as:

John believes that Mary believes that something walks. (12)

$$\text{Bel}[j, \lceil Bel[m, \lceil \exists x. W(x) \rceil] \rceil] \quad (13)$$

For in the latter case the thing walking belongs to the domain of what-John-believes-that-

We shall illustrate these canonical conversion rules for iterated belief sentences for $Bel[j, \ulcorner \neg W(m) \wedge Bel[p, \ulcorner W(m) \urcorner \urcorner]]$:

$$\begin{aligned}
& \|Bel[j, \ulcorner \neg W(m) \wedge Bel[p, \ulcorner W(m) \urcorner \urcorner]]\| = 1 & (14) \\
& \text{iff } \|Bel[j, \ulcorner \neg W(m) \urcorner]\| \text{ and } \|Bel[j, \ulcorner Bel[p, \ulcorner W(m) \urcorner \urcorner]]\| \\
& \text{iff } f_{\mathcal{M}}(j) \models \neg W(m) \text{ and } f_{\mathcal{M}}(\langle j, p \rangle) \models W(m) \\
& \text{iff } \mathcal{M}_j \models \neg W(m) \text{ and } \mathcal{M}_{j,p} \models W(m) \\
& \text{iff } \forall \phi \in \mathcal{M}_j. \|\phi\| = 1 \Rightarrow \|W(m)\| = 0 \text{ and etc.}
\end{aligned}$$

4 Satisfaction for belief sentences

This section should be read as an elaboration of the satisfaction algorithm as defined in the previous chapter.

First of all we need to have a belief space \mathcal{M}_A for every rational agent a we want to reason about. Second, we need to have a special belief space for every arbitrary nesting of belief spaces. So – as was already observed in the above – it is evident that the following sentence requires more than just the belief spaces of John and Mary:

$$\begin{aligned}
& \text{John thinks that Mary thinks she successfully lied to him in} & (15) \\
& \text{saying that it is raining.}
\end{aligned}$$

We will cover these instances of iteration in Prolog by making use of a recursive clause of the satisfaction-predicate for belief-sentences:

```

satisfy (bel(A, P), model(N, D, I), G, Pol):-
    i(A, model(D, I), G, V),
    append(N, [V], NV),
    model(NV, DV, IV),
    satisfy(P, model(NV, DV, IV), G, Pol).

```

A human being is represented by a constant in the language. But when such a person-denoting constant occurs as the first argument to a *Bel*-operator, Prolog picks out the partial model representing a 's belief space. This allows us to reason about human beings as both objects (e.g. "John is heavy") and rational agents (e.g. "John believes that Mary walks"). Both

Mary-beliefs, i.e. $\mathcal{M}_{j,m}$. In the former case, however, the thing walking belongs to the domain of what Mary believes, i.e. \mathcal{M}_m . Therefore we need to determine the assignment to x with respect to D_m (and not with respect to $D_{j,m}$). The expression $W(x)$ must be interpreted with respect to an assignment function $g'[x]g$ that extracts its assignment for x from the domain of bB , i.e. \mathcal{M}_b (and not $\mathcal{M}_{a,b}$).

are related through the function $f_{\mathcal{M}}$ which is implemented in the Prolog code by the call to `i`. We can illustrate this mixed interpretation of persons as in the following sentence:

There is a man who believes George W. Bush to be dead. (16)

The desired meaning to be attributed to this sentence would be:

$$\exists x(Male(x) \wedge Bel(x, Dead(GeorgeWBush))) \quad (17)$$

Our Prolog code is general enough to account for both notions of the human subject to occur within one sentence. The property of masculinity and the assigned belief are naturally related in the meta-language, in which the connection between d_1 within the object-language and $[d_1]$ as a name for the model of a man's belief, is made through $f_{\mathcal{M}}$.

Exercise 4.1. Add the satisfaction conditions for belief-sentences. Add models for various rational agents. A model is a ternary Prolog predicate, such that the first argument is the name of the iterated model, the second argument is the set of entities in the iterated model's domain, the third argument is the interpretation function of the iterated model. A sample set of models would be:

```
model([m], [d1, d2, d3, d4, d5, d6], [
    f(0, jules, d1),
    f(0, vincent, d2),
    f(0, pumpkin, d3),
    f(0, honey_bunny, d4),
    f(0, yolanda, d5),
    f(1, customer, [d1, d2, d5, d6]),
    f(1, robber, [d3, d4]),
    f(2, love, [])]).
model([m,d1], [d1, d2, d3, d4, d5, d6], [
    f(0, jules, d1),
    f(0, vincent, d2),
    f(0, pumpkin, d3),
    f(0, honey_bunny, d4),
    f(0, yolanda, d5),
    f(1, customer, [d1, d2, d5, d6]),
    f(1, robber, [d4]),
    f(2, love, [(d1, d2)])]).
model([m,d2], [d1, d2, d3, d4, d5, d6], [
    f(0, jules, d1),
    f(0, vincent, d2),
    f(0, pumpkin, d3),
    f(0, honey_bunny, d4),
```

```

    f(0, yolanda, d5),
    f(1, customer, []),
    f(1, robber, [d1]),
    f(2, love, [(d1, d2)]).
model([m,d1,d2], [d1, d2, d3, d4, d5, d6], [
    f(0, jules, d1),
    f(0, vincent, d2),
    f(0, pumpkin, d3),
    f(0, honey_bunny, d4),
    f(0, yolanda, d5),
    f(1, customer, []),
    f(1, robber, [d2]),
    f(2, love, [])]).

```

Make sure that as many structurally different belief sentences as possible get interpreted correctly. Especially, observe the way in which nested beliefs and mixed sentences (i.e. those in which a proper noun is used as both an entity and as a model) are handled.

Can you find belief sentences that the present theory is unable to cope with? Explain what the problem is, and whether you think there is a straightforward way of coping with it.