Self-locating beliefs

Eric Johannesson

May 17, 2017

The problem of self-locating beliefs is due to Castaneda (1966), Perry (1979) and Lewis (1979). Let $\mathcal{L}_{\mathbf{B}}$ be a language of first order logic extended by a belief operator and a set of indexical terms. Assume a Kaplan-style semantics, with a relational (Hintikka-style) semantics for the belief operator:

Definition 1 (Frame)**.** A frame is a tuple $\langle W, D, B\rangle$, where W is a non-empty set (of possible worlds), D is a non-empty set (of possible objects) and $B : W \times D \to \mathcal{P}(W)$ is a function assigning a belief set (a set of worlds) to each object in each world (a non-empty set if the object is an epistemic agent in that world, an empty set if it isn't).

Definition 2 (Model)**.** A model of $\mathcal{L}_{\mathbf{B}}$ is a tuple $\langle \mathcal{F}, I\rangle$, where $\mathcal{F} = \langle W, D, B\rangle$ is a centered doxastic frame and I is an interpretation function taking names, indexicals and predicates as argument. If a is a name, then $I(a) \in D$. If i is an indexical, then $I(i) : W \times D \to D$. In particular, if i is the first person pronoun, then $I(i)(w, d) = d$ for all $\langle w, d\rangle \in W \times D$. If P is an n-place predicate, then $I(P) : W \to \mathcal{P}(D^n)$. In particular, if P is the identity predicate, then $I(P)(w) = \{\langle a, b\rangle \in D^2 : a = b\}$ for all $w \in W$.

Definition 3 (Semantics)**.** For any model $\mathcal{M} = \langle W, D, B, I\rangle$, context $c \in W \times D$, world of evaluation $w \in W$ and assignment g,

1. If t is a name, then $[\![t]\!]^c_{\mathcal{M},w,g} = I(t)$.
2. If t is a variable, then $[\![t]\!]^c_{\mathcal{M},w,g} = g(t)$.
3. If t is an indexical, then $[\![t]\!]^c_{\mathcal{M},w,g} = I(t)(c)$.
4. $[\![Pt_1...t_n]\!]^c_{\mathcal{M},w,g} = 1$ iff $\langle [\![t_1]\!]^c_{\mathcal{M},w,g}, ..., [\![t_n]\!]^c_{\mathcal{M},w,g}\rangle \in I(P)(w)$.
5. $[\![\neg\varphi]\!]^c_{\mathcal{M},w,g} = 1$ iff $[\![\varphi]\!]^c_{\mathcal{M},w,g} = 0$.
6. $[\![\varphi \wedge \psi]\!]^c_{\mathcal{M},w,g} = 1$ iff $[\![\varphi]\!]^c_{\mathcal{M},w,g} = 1$ and $[\![\psi]\!]^c_{\mathcal{M},w,g} = 1$.

7. $[\![\forall x\varphi]\!]^c_{\mathcal{M},w,g} = 1$ iff, for all $d \in D$, $[\![\varphi]\!]^c_{\mathcal{M},w,g[d/x]} = 1$.

8. $[\![\mathbf{B}_t\varphi]\!]^c_{\mathcal{M},w,g} = 1$ iff, where $b = [\![t]\!]^c_{\mathcal{M},w,g}$,

   (a) $B(w,b) \neq \emptyset$, and

   (b) for all $w' \in B(w,b)$, $[\![\varphi]\!]^c_{\mathcal{M},w',g} = 1$.

Suppose we want to say, in this language, that

(1) Ralph believes that he (himself) is handsome.

Formally, we would have to say something like

(2) $\mathbf{B}_r Hr$.

However, suppose Ralph suffers from amnesia, so he doesn't believe that he's Ralph anymore. He certainly doesn't believe that *Ralph* (whoever that is) is handsome. If you'd ask him if Ralph is handsome, he'd say he doesn't know. In this situation, it would seem as if (1) is true while (2) is false. So how should we express (1)?

The problem is, we can't. Lewis offers, in my view, the correct diagnosis. The problem is that our models only specifies an agent's beliefs about what the world is like (his location in logical space, as it were), not his beliefs about his location in physical space-time. To illustrate, suppose we have two individuals who are as knowledgeable as they can be about what the world is like. On is on the north pole, the other is on the south pole. In our model, they would be assigned the same singleton belief set containing the actual world only. But one of them believes he's on the north pole, while the other one doesn't. These are distinct doxastic states that cannot be represented in our model.

(P1) If S and S' are in the same doxastic state then, for any φ, S believes that φ iff S' believes that φ.

(P2) S believes that he's on the north pole.

(P3) S' doesn't believe that he's on the north pole.

(C) Hence, S and S' are not in the same doxastic state.

Problem: P1 doesn't seem valid in the first place. Oscar and Twin Oscar are in the same doxastic states (at least on a narrow conception of content). However, Oscar believes that water is wet, but Twin Oscar doesn't.

I agree with Lewis: a belief set needs to be a set of centered worlds, something to indicate where in each world the agent locates himself. Ignoring time, we can let a centered world be a pair consisting of a possible world and an object in the domain of that world. In the simple case of constant domains, the set of centered worlds is thus given by $W \times D$. Again assuming constant domains, we get the following notion of a frame:

Definition 4 (Centered frame)**.** A centered frame is a tuple $\langle W, D, B\rangle$, where W is a non-empty set (of possible worlds), D is a non-empty set (of possible objects) and $B : W \times D \to \mathcal{P}(W \times D)$ is a function assigning a belief set (a set of centered worlds) to each object in each world (a non-empty set if the object is an epistemic agent in that world, an empty set if it isn't).

But there's also a problem at the level of syntax. We need to be able to disambiguate

(3) Ralph believes that he is handsome.

between the reflexive reading

(4) Ralph believes that he (himself) is handsome.

and the indexical reading

(5) Ralph believes that he [pointing at the man in the mirror] is handsome.

Intuitively, as Perry observed, these readings have different truth conditions, even when the man in the mirror happens to be Ralph.

One option is to let the operator $\mathbf{B}$ take a variable as an additional argument, in the following way: if t is a term, x is a variable and φ is a formula, then $\mathbf{B}^x_t\varphi$ is a formula. The idea is to let x, if it occurs in φ, to be interpreted as the reflexive *he/she/it (himself/herself/itself)*. Hence, (4) will be rendered as $\mathbf{B}^x_r Hx$, while (5) will be rendered as $\mathbf{B}^x_r Hi$. This is essentially the solution offered by Castaneda (1966). Let's call the new language $\mathcal{L}^*_{\mathbf{B}}$. As Castaneda (1966, p. 78) points out, the reason we cannot simply introduce a new term t^* to be interpreted has the reflexive pronoun is that we need to distinguish between the following two:

(6) a. Ralph^1 believes that Alf^2 believes that he$_1$ is handsome: $\mathbf{B}^x_r\mathbf{B}^y_a Hx$.
b. Ralph^1 believes that Alf^2 believes that he$_2$ is handsome: $\mathbf{B}^x_r\mathbf{B}^y_a Hy$.

However, the new syntax suggests that *I believe that I am handsome* would have two readings: $\mathbf{B}_i^x Hx$ and $\mathbf{B}_i^x Hi$. In other words, it suggests that the first person pronoun (just like the third person pronoun *he*) can be given both a reflexive *and* an indexical reading in belief reports. Is that plausible? I believe it is. Although in most cases both readings have the same truth value (and the reflexive reading dominates), Maier (2009, p. 272) has a nice example of when they come apart:

> Kaplan is thinking about the time he saw a guy on TV whose pants were on fire without him noticing it (yet). A second later he realized he was watching himself through the surveillance camera system and it was his own pants that were on fire. He reminisces:
>
> (35) I thought I was at a safe distance from the fire
>
> What he thought at the time was 'I am at a safe distance from the fire', which makes (35) true *de se* (i.e. from a first-person perspective). However, the coreferential first-person report construction can also report a third-person *de re* belief that just happens to be about the subject himself:
>
> (36) I thought that I was remarkably calm
>
> The reported thought here may be 'That guy is remarkably calm!' with that guy really referring to Kaplan, the belief subject himself.

The idea is that, while (35) will be true on both readings, (36) will be true only on the indexical reading.

Likewise, the new syntax also suggests an ambiguity in

(7) Ralph believes that he believes that he is handsome.

between the following two:

(8) a. Ralph^1 believes that he$_1^2$ believes that he$_2$ is handsome: $\mathbf{B}_r^x \mathbf{B}_x^y Hy$.
b. Ralph^1 believes that he$_1^2$ believes that he$_1$ is handsome: $\mathbf{B}_r^x \mathbf{B}_x^y Hx$.

To see how the second reading might come about, consider a case in which Ralph is looking at a screen, suspecting he might be the person on the screen.

Ralph thinks the person on the screen looks surprisingly handsome. Normally, Ralph doesn't believe that he's handsome, which is why he's not so sure the person on the screen is him. If the person on the screen *is* Ralph, then Ralph clearly believes of himself that he is handsome. But it would be false (and quite absurd) to say that

(9) Ralph^1 suspects that he$_1^2$ believes that he$_2$ is handsome.

Since evidence for believing *de se* that one is handsome should be easily accessible through introspection, the fact that one does so is not something about which one normally has suspicions. That's why it seems absurd. What's true in the case of Ralph, is rather that

(10) Ralph^1 suspects that he$_1^2$ believes that he$_1$ is handsome.

This goes to show that, even if we supplied a reflexive pronoun for each object in the domain, it would not be enough to disambiguate between (9) and (10).

Definition 5 (Model)**.** A model of $\mathcal{L}^*_{\mathbf{B}}$ is a tuple $\langle \mathcal{F}, I \rangle$, where $\mathcal{F} = \langle W, D, B \rangle$ is a centered doxastic frame and I is an interpretation function taking names, indexicals and predicates as argument. If a is a name, then $I(a) \in D$. If i is an indexical, then $I(i) : W \times D \to D$. In particular, if i is the first person pronoun, then $I(i)(w, d) = d$ for all $\langle w, d \rangle \in W \times D$. If P is an n-place predicate, then $I(P) : W \to \mathcal{P}(D^n)$. In particular, if P is the identity predicate, then $I(P)(w) = \{\langle a, b \rangle \in D^2 : a = b\}$ for all $w \in W$.

Definition 6 (Semantics)**.** For any model $\mathcal{M} = \langle W, D, B, I \rangle$, context $c \in W \times D$, world of evaluation $w \in W$ and assignment g,

1. If t is a name, then $[\![t]\!]^c_{\mathcal{M},w,g} = I(t)$.
2. If t is a variable, then $[\![t]\!]^c_{\mathcal{M},w,g} = g(t)$.
3. If t is an indexical, then $[\![t]\!]^c_{\mathcal{M},w,g} = I(t)(c)$.
4. $[\![Pt_1...t_n]\!]^c_{\mathcal{M},w,g} = 1$ iff $\langle [\![t_1]\!]^c_{\mathcal{M},w,g}, ..., [\![t_n]\!]^c_{\mathcal{M},w,g} \rangle \in I(P)(w)$.
5. $[\![\neg\varphi]\!]^c_{\mathcal{M},w,g} = 1$ iff $[\![\varphi]\!]^c_{\mathcal{M},w,g} = 0$.
6. $[\![\varphi \wedge \psi]\!]^c_{\mathcal{M},w,g} = 1$ iff $[\![\varphi]\!]^c_{\mathcal{M},w,g} = 1$ and $[\![\psi]\!]^c_{\mathcal{M},w,g} = 1$.
7. $[\![\forall x\varphi]\!]^c_{\mathcal{M},w,g} = 1$ iff, for all $d \in D$, $[\![\varphi]\!]^c_{\mathcal{M},w,g[d/x]} = 1$.
8. $[\![\mathbf{B}^x_t\varphi]\!]^c_{\mathcal{M},w,g} = 1$ iff, where $b = [\![t]\!]^c_{\mathcal{M},w,g}$,
   (a) $B(w, b) \neq \emptyset$, and

(b) for all $\langle w', b' \rangle \in B(w, b)$, $[\![\varphi]\!]^{c}_{\mathcal{M},w',g[b'/x]} = 1$.

Definition 7 (Logical equivalence)**.** Two $\mathcal{L}^{*}_{\mathbf{B}}$-sentences φ and ψ are logically equivalent iff, for any model $\mathcal{M} = \langle W, D, B, I \rangle$ and context $\langle w, d \rangle \in W \times D$, $[\![\varphi]\!]^{\langle w,d \rangle}_{\mathcal{M},w} = [\![\psi]\!]^{\langle w,d \rangle}_{\mathcal{M},w}$.

With these definitions in place, we can prove the following:

Fact 1. *$\mathcal{L}^{*}_{\mathbf{B}}$ is more expressive than the fragment of $\mathcal{L}^{*}_{\mathbf{B}}$ where a variable x never occurs freely within the scope of a belief operator $\mathbf{B}^{x}_{t}$. In particular, there's no sentence of the latter kind logically equivalent to $\mathbf{B}^{x}_{i}Px$, where i is the first-person pronoun.*

That means that (the *de se*-reading of) *I believe that I am handsome* cannot be expressed in $\mathcal{L}_{\mathbf{B}}$.

References

Castaneda, H. (1966). 'he': A study in the logic of self-consciousness. *Ratio*, 8(December):130–157.

Kaplan, D. (1989). Demonstratives. In Almog, J., Perry, J., and Wettstein, H., editors, *Themes From Kaplan*, pages 481–563. Oxford University Press.

Lewis, D. (1979). Attitudes de dicto and de se. *Philosophical Review*, 88(4):513–543.

Maier, E. (2009). Proper names and indexicals trigger rigid presuppositions. *Journal of Semantics*, 26(3):253–315.

Perry, J. (1979). The problem of the essential indexical. *Noûs*, 13(December):3–21.