Notes on models of self-locating belief

This is a brief sketch, for those who aren't satisfied with color-coded pictures, of a few details of a formal model, with a little motivation and commentary. The models will use exactly the same abstract objects used in David Lewis's theory of de se belief (centered worlds), to characterize belief states, but will use them in a somewhat different way.

A model is a sextuple $<W, S, T, \geq E, R>$ where

1. W is a nonempty set of possible worlds
2. S is a set of *subjects* or believers
3. T is a set of times
4. $\geq$ is a binary transitive connected anti-symmetric relation on T, a relation that determines a linear order of the times.

T would most naturally have the structure of the points on the real line, but in simple models, we might choose to model only the beliefs of our subjects at certain selected times, so the linear time order might be discrete, and the number of times might be finite. But it is assumed that the ordering is an objective time ordering, and that times can be identified across possible worlds. That is, a certain date (such as Tuesday, April 3, 2007) might be the date on which it rained in Oxford in certain possible worlds, and was sunny there in others.

Two definitions, before characterizing E and R:

a. A *center* is a pair, $<A,t>$, where $A \in S$ and $t \in T$.
b. A *centered world* is a pair $<c, w>$, where c is a center and $w \in W$.

5. E is the set of centered worlds meeting the condition that the subject of the center exist in the world at the time of the center.
6. R is a binary relation on E that is transitive, Euclidean and serial. R must also satisfy an additional condition, which we will state and explain below.

The interpretation of the fifth and sixth elements is this: Subjects may exist at some times at some worlds, and not at others. The set E of centered worlds is restricted to those that are relevant to representing a subject's beliefs at a time in a world. The relation R is the doxastic accessibility relation. To say that $<<A,t>, x>R<<B,t^*>, y>$ is to say that it is compatible with what A believes at time t in world x that she is in world y, that she is person B, and that the time is time t*.

Given R, each centered world in E determines a set of centered worlds – those that are R-related to it. Call a pair consisting of a centered world and its R-related set a *belief state*, and call the determining centered world the *base (centered) world*, and the determined set the *belief set.* The role of the center of the base world is to specify the person whose beliefs are being represented, and the time at which she has those beliefs. The role of the centers of the centered worlds in the belief set is to represent where that subject takes herself to be in the world that, for all she believes, is actual. If Alice thinks, on Sunday, that it might be Monday, and that she might be Clara, rather than Alice, then a world centered on Clara on Monday will be compatible with what she believes (on Sunday).

We impose the following condition on the relation R:

> **(*) For any centers, c, c' and c*, and worlds w and x: if <c, w>R<c', x> and <c, w>R<c*, x>, then c' = c*.**

What this condition requires, intuitively, is that ignorance or uncertainty about where one is in the world is always also ignorance or uncertainty about what world one is in. Even in the highly artificial case where a subject believes that he will, in the actual world, be in two qualitatively indistinguishable situations at different times, t_1 and t_2, without knowing which time it is, it will remain true that (as he would put it at the time) the world where *this* token thought is occurring at time t_1 (and where another like it will occur at t_2) is a different (uncentered) possible world from the possible world in which *this* (token) thought is occurring at time t_2 (and another like it occurred at t_1).[1]

This crucial condition is the main point at which the proposed model differs from Lewis's account of de se belief, which allows that a case of ignorance might be represented by two centered worlds – two "predicaments", to use Adam Elga's term – centered at different points within the same world. So, for example, we reject Lewis's representation of the beliefs of the notorious two gods who (according to Lewis's account) are each omniscient about what world they are in, but ignorant of which of the two gods he is. We say, instead, that there are two qualitatively indiscernible worlds, and that neither god knows which of the two is actual. By requiring that ignorance and doubt always be represented by distinctions between possible states of the world, we allow for the calibration of the states of belief of different believers, and of a believer at different times. Even though belief *states* are represented by sets of centered possible worlds, the *contents* of belief can be taken to be ordinary propositions – sets of uncentered possible worlds. So in the interpretation of statements of the form "x believes that φ", the "that φ" will denote a set of (uncentered) possible worlds, even though the centers determined by a particular belief state may play a role in determining which proposition is denoted by a that-clause with indexical expressions in it. By taking the contents of belief to be (uncentered) propositions, we can straightforwardly compare the beliefs of different subjects, and we can model the way assertions change the context in a straightforward way. We can also model the dynamics of belief for a single agent – the facts about preservation and change of belief – in a straightforward way. In particular, we can apply a standard belief revision theory to a rational subject with a prior belief state at time t, who then receives some new information at time t* while remembering her prior state. Even if some of her prior and posterior information is self-locating (suppose, for

[1] I don't want to rest anything on the assumption that the same token thought might have occurred on a different day. It might be a counterpart token that occurred, in the other possible world, on the other day. What one needs to motivate the assumption that there are two (uncentered) possible worlds here is just that the thought that takes place, in the actual world, at the other time is a different token thought

example, she didn't know what time it was at t, or how much time passed between t and t*), she can still revise her beliefs in the standard way. If we want to add to our model probability measures on belief states to represent degrees of belief, this will be as straightforward as in standard belief logic models, and we could then represent the assumption that rational subjects will revise by conditionalization.

In the standard Hintikka-style semantics for logics of knowledge and belief, ordinary uncentered possible worlds are the relata of the doxastic or epistemic accessibility relation. The identity of the believer, and (implicitly) the time of belief are built into the relation. In a theory of this standard kind with multiple believers, there will be multiple accessibility relations, one for each believer. Our models, in contrast, need only a single doxastic accessibility relation, since the identity of the believer and the time of the belief are determined by the center of the first relatum. By putting the believer and the time of belief into the relata, rather than the relation, we not only provide the resources to represent self-locating belief, but also a more flexible framework for representing the relations between the beliefs of different believers, and of a single believer at different times.

In the standard belief semantics, the representation of iterated belief is a simple matter: If A and B are two believers, and R_A and R_B are their doxastic accessibility relations, then it will be true, in world w, that A believes that B believes that φ iff for all worlds x such that wR_Ax and all worlds y such that xR_By, φ is true in y. One can define the set of possibilities compatible with the *common beliefs* of A and B in terms of the transitive closure of the two relations R_A and R_B, or more generally, the common beliefs of a set of subjects in terms of the transitive closure of the set of accessibility relations for the subjects in the set. In our models, the representation of iterated belief is a little more complicated, but the complications reflect complexities in the phenomena being modeled, and the increased flexibility in the representational resources of the model. The first complication come from the fact that we have made explicit that belief is relative to time, something that is ignored in the standard theory. One might represent A's beliefs at t about what B believes at some different time t', but let's ignore that for now, and just focus on A's beliefs at some time t about what B believes at the same time. Still, A may not know what time it is, so the actual time at which A has her beliefs may be different from the time she takes it to be. For example, if A mistakenly believes on Tuesday that it is Monday, then there will be a difference between "A believes (on Tuesday) that B *now* believes that φ" and "A believes (on Tuesday) that B believes *on Tuesday* that φ." The truth of the former will depend on what B believes *on Monday* in the worlds compatible with A's beliefs, while the latter will depend on what B believes *on Tuesday* in those same worlds.

A second complication is this: Because of the intentionality of belief, A may have different beliefs about B's beliefs, relative to different ways of thinking about him. Suppose I am sitting in the bar with a man in a brown hat who is in fact Ortcutt, but I am not sure whether he is Ortcutt or O'Leary. We are watching the Red Sox on the television, and I believe that the man in the brown hat believes that the Red Sox are losing, since they *are* losing, and it is evident that the man is paying attention to the game. But I am not sure whether *Ortcutt* believes this, since for all I know, the man at the bar is O'Leary, and Ortcutt is somewhere else, blissfully ignorant of the state of the game.

In the simple case, where it is assumed that A knows who B is, we can ignore this, but for the general case, we need to relativize iterated belief, (what A believes about what B believes), to a way that A thinks of B (dare I call it a mode of presentation?), formally represented by a function from worlds to individuals. A function of this kind will represent (in a given world) a possible way of thinking *about B* if it takes B as its value in that world. In the simple case (where A knows who B is), this function will be the constant function, taking B for all arguments, but in the general case, it might be a variable, or non-rigid, individual concept. (We can assume that the function is everywhere defined within the worlds in A's belief set, since we can assume that if A identifies B as "the F", then A believes that there is a unique F.)

To use nonrigid functions, or individual concepts, to characterize the centered worlds does not add any new centered worlds to our model: it just gives us new ways to generalize about them. Suppose f is the nonrigid function, or individual concept, expressed by "the man in the brown hat", and that f(w) = Ortcutt. Then the centered world <<f, t>, w> is just the centered world <<Ortcutt,t>, w>. But when we quantify over centered worlds, f may take different values for different values of w. For example, consider this generalization:

> For all worlds x and y and for all subjects C, if <<A,t>, w>R<<A,t>,x>,
>
> and <<f, t>, x>R<<C, t>, y), then y ∈ φ.

This says that in world w, A believes that the man in the brown hat believes that φ.

Once we have a clear account of iterated belief, we can use it to define a notion of *common belief* for a group of individuals at a given time and the properties of a common belief state will be generated by the iterative process. It is common belief (among the members of group G) that φ iff all believe that φ, all believe that all believe that φ, all believe that all believe . . . , etc. To keep things simple, we might assume that everyone in the group knows who everyone else in the group is, but we can also model cases where the members of a group have some common way of identifying each other, even though they may not know who the others, or even themselves, are. So, for example, we might model the common ground (presumed common beliefs) of a conversation between two amnesiacs trying to figure out who and where they are, and what time it is, by pooling the meager information that they each have. In general, the common ground that is determined by the iterative process will generate a representation that parallels the representation of an individual belief state; it will have the same structure, but with centered worlds with multiple individuals at their centers. An individual belief state is a pair consisting of a centered world (the base world) and a set of centered worlds (the belief set). The common ground can also be represented by a base world and a common belief set, but with a sequence of individuals (all those in the relevant group) at the centers instead of a single individual. The sequences of individuals at the centers of the common belief worlds will represent where the members of the group mutually locate themselves and each other in the possible worlds compatible with their common beliefs.

Our models have an accessibility relation only for belief, but a subject might also have other self-locating attitudes. In some cases, self-location in possible worlds that are not compatible with the subject's belief is derivative from self-location in belief-worlds. (This fact was exploited in the account of the Sleeping Beauty case.) Suppose I don't know whether I am A or B, but I do know that if the coin had landed heads (which I know it did not), then I would have won the bet. What I know is that if I am in fact A, then if the coin had landed heads, A would have won, and if I am in fact B, then if the coin had landed heads B would have won the bet. A second kind of case of derivative self-location in worlds incompatible with the subject's beliefs is iterated belief. To represent my belief that John believes that I am a plumber, I need to locate myself in the possible worlds that, for all I believe, are compatible with what John believes. This self-location is derivative in the same way as in the case of the counterfactuals.