Attitudes and Ascriptions in Stalnaker Models

Abstract

What role, if any, should centered possible worlds play in characterizing the attitudes? Lewis (1979) argued (in effect) that, in order to account for the phenomena of self-location (Perry 1977, 1979), the contents of the attitudes should be taken to be centered propositions (i.e. sets of centered worlds). Stalnaker (2008, 2011, 2014), however, has argued that while centered worlds are needed to characterize e.g. belief *states*, the *contents* of such states should be understood as ordinary, uncentered propositions (cf. Hintikka 1962). But Stalnaker does not, as is common, provide a semantics of attitude ascriptions based on the models he develops of the attitudinal states themselves. This paper explores the prospects for doing so. It argues that a simple Millian semantics does not yield the principles of knowledge and belief Stalnaker endorses; and that a retreat to descriptivism brings with it problems of its own. A technical appendix contains novel and pertinent results in doxastic/epistemic logic.

Keywords

Self-Location, Centered Worlds, Propositional Attitudes, Attitude Ascriptions, Stalnaker Models, Factivity, Transparency, Introspection, Doxastic Logic, Epistemic Logic

1. Introduction

What role, if any, should centered possible worlds play in characterizing the attitudes? Lewis (1979) argued that, in order to account for the phenomena of self-location, the contents of the attitudes should be taken to be (what are in effect[1]) centered propositions, or sets of centered worlds. More recently, however, Stalnaker (2008, 2011, 2014) has suggested that while centered worlds are needed to characterize e.g. belief *states*, the *contents* of such states should be understood as ordinary, uncentered propositions - just as in the tradition of doxastic and epistemic logic inspired by Hintikka (1962).

The phenomena of self-location are, by now, quite familiar.[2] Nevertheless, by way of illustration, allow me to briefly outline Perry's (1977) case of Lingens, the amnesiac lost in the Stanford library. Lingens reads a biography of himself, and so knows many things about Lingens; moreover, he reads a description of the Stanford library, and thereby learns many things about it. But Perry's thought was that Lingens may nevertheless fail to know these various things about himself and his surroundings, *considered as such*, until such time as he is in a position to say, '*I* am Lingens, and *this* is the Stanford library' – that is, until he

[1] In fact, Lewis suggests that the content of an attitude is a property, which is to say 'the set of exactly those possible beings, actual or not, that have the property in question' (1979: 515). The possible beings at issue, however, are time-bound, world-bound individuals, and so – if a center is a pair of an individual and a time, and a centered world is a pair of a center and a world - Lewis' properties are formally equivalent to sets of centered worlds, i.e. centered propositions. Notice that the inclusion of an individual, assuming it to be material, tacitly brings with it a spatial location, namely the location occupied by that individual in the relevant world at the time in question. Below, we will simplify the discussion by ignoring times, and taking centers to be simply individuals.

[2] Perry has presented a number of cases in which a subject fails, in some (disputed) sense, to self-locate: for instance, his (1979) sugar shopper, who is making a mess, but fails to identify himself as the mess maker; his (1979) mountain walker, who knows the best route from where he in fact is to where he wants to go, but fails to take it because he is lost (i.e. doesn't know where he is); Heimson, who erroneously thinks he's Hume (1977); and the (1977) case of Lingens, described below.

is able to locate himself (in this sense) amongst the various people and places that there are; and he may accordingly be unable to act appropriately - say, by giving his name to the librarian when checking out a book, or by walking to the cafeteria when lunchtime comes.

Lewis (1979) discussed this case, before introducing his own case of the two gods, both of whom know all of the (ordinary, uncentered) propositions that are true of the (uncentered) possible world they inhabit, but who nevertheless fail to know which god they are. These gods represent an extreme, limiting case of the failure to self-locate, or identify: their ignorance is *irreducibly* self-regarding, or *de se*; and it was to account for such ignorance, and the knowledge that is obtained from eliminating it, that Lewis introduced centered propositions to act as the contents of attitudes. The idea is that it is consistent with what each of the gods knows, (only) that they are centered on themselves in the world they both inhabit, and that they are centered on the other god in that same world: thus, only one uncentered world is consistent with their knowledge – this is what makes them propositionally omniscient; but there are two different centers consistent with their knowledge – and it is this which makes them unable to self-locate (i.e. to locate or identify themselves within that world), and which constitutes their ignorance.

While recognizing the force of examples like Perry's, Stalnaker (2008) thinks Lewis went too far: any ignorance regarding self-location must, on Stalnaker's view, be ignorance regarding the truth-value of some ordinary, uncentered proposition; there is, he suggests, no knowledge (or belief) that is irreducibly *de se*, and Lewis' gods are impossible. Moreover, Stalnaker (2011) worries that if the objects of the attitudes were centered propositions, communication would be impossible: one could not assert the content of one's attitude, and have one's audience understand, accept, and thereby come to stand in the same relation to, the very same content. Some have responded to this objection by showing how communication might work even if the contents of the attitudes were centered propositions: thus, Egan (2007) and Moss (2012) have defended de-centering accounts on which what one asserts (an ordinary proposition) is not what one believes (a suitably related centered proposition), though what one's hearer understands is what one asserts; while Ninan (2010) has defended a re-centering account on which what one asserts (when one is sincere!) is what one believes (some centered proposition), but, in understanding, one's hearer does not entertain this very same thing (but instead grasps some other related centered proposition).[3] With this debate ongoing, I do not take Stalnaker's objection from communication to be decisive against Lewis' account: but I find the thought that Lewis' two gods are impossible to be intuitively plausible; and in any case, Stalnaker (2008, 2014) develops his own intriguing account of the attitudinal states involved in cases of self-location, in which centered worlds play a distinctive new role – an account which is worthy of consideration in its own right.

The remainder of this paper is devoted to the investigation, development, and criticism, of Stalnaker's positive proposal. In particular, while Stalnaker devises models, employing centered worlds, of attitudinal states, he does not (as is usual) provide a semantics for a language capable of ascribing the attitudes modeled. This paper begins to address this lacuna; and it draws out some difficulties for Stalnaker's view.

2. Millian Semantics in Stalnaker Models

I begin by describing Stalnaker's models and showing how they can, on the face of it, be used to account for the case of Lingens discussed above. I then give a syntactic description of languages containing attitude

[3] For what it's worth, Frege (1918) seems to have opted for something like the de-centering account – Lauben, despite thinking of himself in the way that only he can, uses 'I' with the sense of *he who is speaking to you now*.

ascriptions, and suggest a simple, Millian semantics which vindicates this first appearance; but I show that on this approach certain key principles of doxastic logic endorsed by Stalnaker (2006) are not validated.

2.1 Stalnaker Models

As we have seen, Stalnaker thinks there is a role for centered worlds to play in characterizing the attitudes subjects have in cases of self-location. Accordingly, he devises a class of models of belief states – and, one might think, by extension, knowledge states - comprising a set W of worlds, a set S of subjects, a set E of centered worlds (built – as pairs - from the subjects and the worlds), and an accessibility relation R on E.[4] The accessibility relation R in these models is subject to a number of constraints: it must be serial - reflexive, one assumes, in models of knowledge - transitive and Euclidean;[5] and it must also respect what I will call *Stalnaker's constraint*, which says that for all worlds w and x, and all centers c1, c2, and c3, if <c1,w> R <c2,x> and <c1,w> R <c3,x>, then c2=c3. I will discuss (both variants of) the first three of these constraints below, but it is this last which ensures that the objects of the attitudes are propositional: uncertainty about who (or, more generally, where or when) you are requires uncertainty about which world you are in; thus, contrapositively, belief about and/or knowledge of which world you are in brings with it belief regarding, and/or knowledge of, who (where, when) you are – and so, to repeat, Lewis' gods are impossible.

On the face of it, we can capture both the knowledge and the ignorance of Lingens, lost in the library, in a model of this kind. In particular, suppose that after reading for some time, Lingens, still amnesiac, begins to feel hungry. Perhaps at first he experiences this merely as some form of discomfort or other, but eventually he comes to recognize it as hunger. Then it seems that he thereby comes to believe, and indeed know, that he is hungry; yet he doesn't know, or even believe, that Lingens is hungry, since, for all he knows or believes, Lingens is elsewhere, recently fed! And this would appear to be captured in a simple Stalnaker model comprising just two worlds, @ and w, and two subjects, Lingens and Perry, with the following accessibility relation on the set of centered worlds (which itself comprises all four combinations of the subjects and the worlds): R relates every centered world to itself (so that it is reflexive), plus R relates Lingens at @ to Perry at w and vice versa. No other centered worlds are R related. (See appendix.) Then, assuming that Lingens is hungry in @ and Perry is hungry in w, yet Lingens is not hungry in w, it seems Lingens should be able to say truly, in @, both 'I know I am hungry' and 'I don't know Lingens is hungry' (and similarly in the case of 'believe'). The trick, however, is to devise a semantic theory for a language comprising attitude ascriptions that makes the relevant predictions.

2.2 Millian Semantics for Attitude Ascriptions

Syntactically, the languages we will consider contain terms – both names, and a special (logical) constant, 'i'; n-place predicates for each n (including 0); truth-functional connectives; and an attitude verb - either 'believes' or 'knows' - taking a term and a sentence to form a sentence. For simplicity, there are no variables or quantifiers.

Semantically, we extend Stalnaker models by adding an interpretation function I; and we define denotation and truth, recursively, relative to a centered world. Thus, relative to a given centered world, the logical constant 'i' denotes the subject at its center; and names are given the Millian treatment, so

[4] I simplify here by ignoring times.

[5] R is: serial iff $\forall x \exists y Rxy$; reflexive iff $\forall x Rxx$; transitive iff $\forall x \forall yz(Rxy \cdot Ryz \rightarrow Rxz)$; and Euclidean iff $\forall x \forall y \forall z(Rxy \cdot Rxz \rightarrow Ryz)$.

that, relative to any centered world they denote whatever element of S the interpretation function assigns to them. Predicates are assigned (ordinary) intensions, i.e., functions from (uncentered) worlds to subsets of S^n (or truth-values in the case of 0-place predicates). Atomic sentences are true relative to a centered world iff the denotations of the terms at that centered world, taken in order, are in the extension of the predicate at the (uncentered) world (or, in the case of 0-place predicates, just in case their extension is truth at that world). The truth-functional connectives are given the expected, standard treatment; and the clause for the attitude verb says that $S\ believes/knows\ \varphi$ is true, relative to <c,w> iff φ is true at <c',w'> for all centers c' and worlds w' such that <s,w> R <c',w'>, where s is the denotation of *S* at <c,w>.

Given this semantics, if we render 'I' by 'i', and the other expressions in the obvious ways, 'I believe/know I am hungry' comes out true relative to Lingens and actuality in the model of the previous section (supplemented with an appropriate interpretation function), while 'I believe/know that Lingens is hungry' comes out false relative to that same centered world in that model – exactly as desired. (See appendix.) Despite this initial success, however, the semantics provided above is not entirely unproblematic.

2.3 Principles of Doxastic and Epistemic Logic in Hintikka Models

Stalnaker (2006) has argued that the correct logics of belief and knowledge validate certain key doxastic and epistemic principles. The first of these that is relevant for our purposes[6] is:

(D) $If\ S\ believes\ \varphi, then\ S\ does\ not\ believe\ not\ \varphi.$

This is a consistency principle for the (perhaps idealized[7]) notion of belief; and, in a standard (Millian) semantics based on (what we can call) Hintikka models (following Stalnaker, 2014: 10), it is validated if, and only if, models are required to contain an accessibility relation (for each subject, and now on the set W of uncentered worlds) that is serial. The idea is that if a subject's beliefs at w are consistent, as D requires, then they are possibly true; which is just to say that some possible world x is consistent with what that subject believes in w, and so x is doxastically accessible for her in w.

Similarly, the factivity of knowledge is widely held to be expressed by:

(T) $If\ S\ knows\ \varphi, then\ \varphi.$

And again, on the standard approach based on Hintikka models, this is validated if, and only if, the accessibility relation is required to be reflexive: here the idea is that what a subject knows at a world w cannot be incompatible with what obtains at w; for knowledge (at w) entails truth (at w). Notice that reflexivity entails seriality, just as truth entails consistency. T is a strengthening of D.

Next, consider:

(4) $If\ S\ believes/knows\ \varphi, then\ S\ believes/knows\ S\ believes/knows\ \varphi$; and

[6] Both of Stalnaker's preferred logics – the modal logic KD45 in the case of belief, and S4.2 in the case of knowledge - also validate the rule of necessitation, which says that if φ is a theorem, so is *S believes/knows* φ, and the principle K, which says that *if S believes/knows that if* φ *then* ψ*, then if S believes/knows that* φ *then believes/knows that* ψ*.* These assumptions are unproblematic if we idealize in such a way as to interpret *S believes/knows that* φ to mean that the proposition that φ follows from what S believes/knows. Given the discussion below, however, it is not clear that this idealization suffices for Stalnaker's purposes (cf. Stalnaker, 2010: 232). Thanks to EU and JW for discussion.

[7] See previous note.

(5) $If\ S\ does\ not\ believe/know\ \varphi, then\ S\ believes/knows\ S\ does\ not\ believe/know\ \varphi.$

These are often thought of as transparency principles – those of positive and negative introspection respectively - for the attitudes (cf. Stalnaker, 2014: 232); and on the standard approach, they are validated if, and only if, the models are required to have an accessibility relation that is transitive and Euclidean respectively.[8]

2.4 Counterexamples in Stalnaker Models

It should therefore come as something of a surprise to discover that there are counterexamples to both 4 and 5 given the above semantics of attitude ascriptions based on Stalnaker models – despite the fact that the accessibility relation is required to be both transitive and Euclidean. For instance, in the model sketched above, 'Lingens doesn't know Lingens is not hungry' is true (relative to Lingens and actuality), but 'Lingens knows Lingens doesn't know Lingens is not hungry' is false (relative to that same centered world), so that the corresponding instance of 5 fails. The reason, of course, is that Lingens is hungry in @ in that model, so doesn't know Lingens is not hungry, relative to himself and @. However, it is consistent with what he knows at (himself and) @ that he is at x, where the person he might be, for all he knows, is (the amnesiac) Perry. Yet at x – which is consistent with Lingens' knowledge at (Lingens and) @ - *Lingens* knows exactly (who he is - namely, Lingens – and) which world he's in - namely, x. Moreover, Lingens is not hungry in x. So, relative to Lingens and x, Lingens knows Lingens is not hungry. So it is consistent with what Lingens knows relative to himself and @ that Lingens knows Lingens is not hungry, and 5 fails. (See appendix for further details.) Similarly, in a simple three-world model we can have 'Lingens knows Lingens is hungry' come out true (relative to a centered world) while 'Lingens knows Lingens knows Lingens is hungry' comes out false (relative to that same centered world), thereby invalidating 4. (See appendix for proof.) Finally, and perhaps most astonishingly, on the semantics given, T is not valid in reflexive Stalnaker models! (See below, and appendix, for details.)

Prima facie, these results are bad for Stalnaker: after all, he has committed to the transparency of many of the attitudes;[9] and the factivity of knowledge must be one of the least contentious principles in all of philosophy! But they are also just plain puzzling. Why should it be that the various principles fail in Stalnaker models, even when the accessibility relation has the formal properties that suffice to validate them in Hintikka models? In what follows, I consider how Stalnaker's preferred principles might be salvaged, as well as what happens if they are abandoned.

3. Retreat to Descriptivism

How, if at all, can Stalnaker uphold the principles of belief and knowledge he has advanced? It emerges that he can employ a descriptivist strategy - which, I suggest, encounters some difficulties of its own.

[8] Stalnaker's preferred logic of belief, KD45, validates both 4 and 5; whereas his preferred logic of knowledge, S4.2, validates 4 but not 5. Nevertheless, as Stalnaker (2006) notes, many have (erroneously) considered S5 to be the correct logic of knowledge, and for this reason I have formulated both 4 and 5 using both 'believes' and 'knows'.

[9] Stalnaker endorses positive introspection, not only for belief and knowledge, but also for acceptance (2014: 232); and he also endorses negative introspection for acceptance and belief (2014: 232), though not in the case of knowledge (2006). He has also insisted that the (Hintikka) accessibility relation for presupposition is serial, transitive, and Euclidean (1999: 99), suggesting that he accepts both transparency principles for that attitude as well.

3.1 Saving Factivity?

I begin by reconsidering the failure of the T schema. First, it might be thought that this is a problem, not for Stalnaker's models of belief (and structurally similar attitudes, such as that of acceptance[10]), but for my proposed extension of Stalnaker's framework to cover the case of knowledge. Yet while this is strictly speaking true, it is, I think, to be expected that Stalnaker's models should enjoy the same generality as the more standard Hintikka models: thus, the proposed extension seems to me to be desirable; we should not rest content with the ability to model just some of the attitudes. This first response fails.

A second response looks in more detail at the nature of the counterexamples. For example, the sentence 'Lingens knows i is hungry' is true at <Perry, @> in the first (two-world) model described above, but 'i is hungry' is not true relative to these same parameters (see appendix). The reason, of course, is that in the first sentence 'i' picks out Lingens, while in the second it designates Perry: accordingly, the proposition expressed by 'i is hungry' is different when it is embedded under 'Lingens knows' than when it occurs unembedded; thus, no proposition is said to be both known by someone and yet untrue.[11] The problem therefore does not arise in a language (perhaps the one Stalnaker intended) whose atomic sentences are all propositional variables – and if we enhance such a language with propositional quantifiers, we can express factivity not with the T schema, but directly as the claim that all known propositions are true. The problem therefore resides in the semantics, it might be said, and not the models on which they are built. In particular, 'i' is what Kaplan (1989) calls a monster: the attitude constructions shift a feature of the context (namely the center[12]); and it is this combination which, in our example, leads to the failure of T.

This is, of course, all true. But notice that: (a) Stalnaker himself explicitly endorses the existence of monsters (2014: 210-216), so it is not clear that he can mount this defense; (b) the fact that 'i' is a monster was crucial to capturing the intuitions in our original case of Lingens in the library; and perhaps most importantly, (c) the other counterexamples (to 4 and 5) don't go away in the language whose sole atomic sentences are propositional variables. The problem has not yet been (dis)solved.

3.2 The Narcissistic Language

Consider a language that is like the Millian ones above, except that it lacks names: thus, its sole singular term is 'i'; and accordingly, speakers of the language can only speak about themselves.[13] In this narcissistic language there are no counterexamples to the logical principles discussed above in Stalnaker models having the corresponding formal features (see appendix for proofs). (This provides further evidence that it is not the treatment of 'i' that is to blame for generating the counterexample to T above.) In fact, thanks to Stalnaker's constraint, the accessibility relation (in a given model) yields, for each subject, at each world, an 'individual concept' (2008: 73), or function from (ordinary, uncentered) worlds to individuals, which captures that subject's mode of self-presentation (i.e., roughly, her way of thinking about herself

[10] See previous note.

[11] Roughly, the counterexample has Perry saying falsely, 'If Lingens knows he is hungry, then I am hungry'. Big deal!

[12] The assumption that the center is a feature of the Kaplanian context, and not of the circumstance of evaluation, is crucial for the diagnosis that 'i' is a monster. This assumption might be abandoned; but then it is hard to make sense of the notion of truth at a centered world – especially in such a way as to respect the Stalnakerian thought that truth-evaluable contents are sets of uncentered worlds. (Notice that the world component of a centered world plays something of a double role, both as context and as circumstance. In an earlier draft, I separated out these two functions, construction a two-dimensional semantics: but the resulting complexity was not exploited in the language; thus, the virtues of simplicity prevailed. Thanks to OM for discussion on this point.)

[13] In the actual world. What they say might concern other individuals relative to other possible worlds.

as herself) at that world. In the above semantics, the term 'i' is interpreted, relative to a given centered world, as expressing just such an 'I-concept' (2014: 120) of that subject at that world. In short, the semantics for 'i' exploits the structure of the model, and is entirely appropriate to it.

Crucially, then, while the names in the Millian language are rigid designators, 'i' expresses a non-rigid individual concept (of the kind expressed by ordinary definite descriptions); and it is this difference which accounts for the validity of the logical principles in the narcissistic language despite their failure once names are introduced. Thus, Stalnaker can appeal to a kind of *descriptivism* to defend his position. What he must maintain, specifically, is that, for each subject, there is an individual concept relative to which (knowledge is factive and) her attitudes are transparent. This is, of course, a contentious claim. But either it is true, or it isn't. I consider each possibility in turn.

Let us grant that there are such individual concepts. Nevertheless, one concern arises already within the narcissistic language. Frege says that 'everyone is presented to himself in a particular and primitive way, *in which he is presented to no-one else*' (1956: 298, my emphasis). Now suppose that this is true of the individual concept with which Lingens, for example, has transparent access to his own beliefs. If he then says to Perry, 'If i believes i is hungry, then i believes i believes i is hungry', he may employ this self-concept; but Perry may not be able to grasp the thought in question, even though it is an ordinary, uncentered proposition. Communication risks becoming impossible in practice, even if not in theory.

Of course, we need not accept Frege's account of a subject's *self* concept (or apply it in the case at hand). But Stalnaker does seem to hold that one can only grasp 'now' thoughts at the time in question (2008: 85), so it is perhaps not too much of a stretch to think the same might be true of 'i' thoughts on his view. Indeed, Stalnaker maintains that in some cases a subject might help to locate herself in part by appeal to '*this* token experience' (2008: 61) she is undergoing, which is presumably not a way in which another subject might identify her![14] The risk is therefore real: the narcissistic language may ultimately be a private language.

3.3 Third Person Ascriptions and Common Attitudes

We cannot, of course, rest content with the ability to speak only about (and to!) ourselves. Indeed, it is perhaps Stalnaker who, most of all, needs to be able to engage in third-person ascription of attitudes. 'The most important concept of the pragmatic framework that I have used for many years', he says, 'is the concept of *common ground*' (2014: 2), where it is common ground that φ in a group G, according to Stalnaker, if and only if each member of G accepts φ, accepts that each accepts φ, and so on, *ad infinitum* (cf. the definition of common belief in Stalnaker, 2008: 73). Thus, it is crucial for Stalnaker to have an account of iterated and interpersonal attitude ascriptions in the third person.

Suppose Perry wants to report the fact expressed by Lingens above. He might say, 'If Lingens believes that he's hungry, then Lingens *believes* Lingens believes he's hungry' (with a little emphasis on the italicized

[14] In this respect, Stalnaker's approach calls to mind that of Russell (1910), on which, for example, the friends of Bismarck used the name 'Bismarck' to express some such description (or individual concept) as *the cause of these sense data*. (As we shall see below, I don't think this is a mere coincidence.) But sense data are - roughly speaking, at least - (constitutively) internal to subjects, and so are token experiences. Thus, the current claim is in tension with Stalnaker's professed aim in his (2008) to reconcile the perspectivalism required to do justice to the phenomena of self-location, on the one hand, and on the other, content externalism, i.e., the thesis that 'propositional content is the kind of thing that can be characterized only in terms of materials that are 'external to the mind' of the subject whose thought has that content' (2008: 111).

instance of 'believes'). But if Perry is to mean what Lingens does, the words 'Lingens' and 'he' must express Lingens' *self* concept (in the actual world).[15] Of course, there is some danger, as we have seen, that it is not possible for Perry to grasp this concept. But if that's right, one would not expect him to be able to express it either. So one consequence of adapting the descriptivist strategy employed in connection with the narcissistic language to the case of third-personal ascriptions is that we extend the range of people who must be able to grasp the individual concept relative to which a subject's attitudes are transparent: not only the subject, and the hearer, but also the speaker, must think of the subject under the relevant mode of presentation.

Moreover, when it comes to the common attitudes – such as common knowledge, or common acceptance - within a given group, it is not only the subject, as well as the ascriber of a particular individual attitude, and his or her audience, but also the other group members, who must share the mode of presentation of the subject in order to ensure transparency. This might be thought to demand too much of the individual concept in question: it must be both such as to provide, for the subject, something like immunity to error, and even agnosticism, through misidentification (Shoemaker, 1968) about his or her attitudes, and also, at the same time, (more or less fully) publicly available. Are there really such MOPs?

Of course, if the individual attitudes in question are not transparent, or at least subject to positive introspection, then almost nothing is common knowledge, or common belief, etc.; in which case the context set, for example – defined as the set of worlds that are compatible with the common ground - will be so vast as to not realistically represent the live options available in a given conversation. But in light of the above considerations, this means that, for Stalnaker, any interesting notion of a common attitude will not be reducible to that of the individual attitudes *simpliciter*; rather, it reduces to the individual attitudes *plus* (or *relative to*) the group's collection of ways of thinking of themselves (not each individually as him or herself but collectively) as themselves. This may not be problematic, but again, it is contentious. In any case, Stalnaker seems to recognize this irreducibility when he suggests that the centers in his models must be generalized to sequences of individuals in order to capture the common attitudes (2008: 73-74). Models of the individual attitudes of group members alone do not predict the common attitudes of the group on Stalnaker's approach.

3.5 Abandoning Transparency

Given the difficulties involved in securing transparency, it seems worth asking what happens if it is abandoned (and with it the assumption that there are individual concepts serving to secure it). In particular, suppose that we do not require the accessibility relation to be transitive in Stalnaker models. (We would like, in any case, and for reasons of generality, to be able to model such attitudes!) Then to ensure that it determines an individual concept, at each world, for each subject, we must impose Stalnaker's constraint not on R itself, but on its transitive closure R*. Still, if we use this individual concept to interpret names (and pronouns), as we did above, we will face a dilemma: either the language will fail to be productive; or it will be unlearnable.

[15] Formally, we might capture this by dropping 'i' from the language, and replacing it with a series of variables, co-indexed to the names (which might be thought of as binding them). We can then assign an individual concept (perhaps subject to certain restrictions) to each index. (After all, Stalnaker says 'we need to relativize' 2008: 72) our ascriptions to individual concepts.) As a side effect we might hope to thereby eliminate the problematic instances of T from the language: when the index on the embedded term doesn't match that on the unembedded term the sentence as a whole won't count as an instance of T.

To see this, note that we may think of the individual concept in question as stitched together from a series of partial functions from (uncentered) worlds to individuals: the first is defined on the set of worlds that result from dropping the center of some accessible centered world; the second is defined on the set of worlds we get if we drop the center of some accessibly accessible centered world; and so on. But then, the first partial function serves to interpret e.g. 'Lingens' when it is embedded under 'Lingens believes that'; the second when it is embedded under 'Lingens believes that Lingens believes that'; and so on. Yet if transitivity fails, these various domains are distinct: and so the first of these functions does not, in general, determine the second; and thus, the meaning of the name (on its second occurrence) in 'Lingens believes that Lingens is hungry' does not determine its meaning (on its third occurrence) in 'Lingens believes that Lingens believes that Lingens is hungry'. The language is not (semantically) productive.

Alternatively, if we think of the meaning of the name as given by the *total* function that results from stitching together the various partial functions, then there will be no reason to think that we can learn this meaning. For instance, on this proposal, 'Lingens' doesn't just mean 'whomever Lingens might be, for all he knows'; for when it is doubly embedded under 'Lingens knows that' it means something closer to 'whomever the subject whom Lingens might be, for all Lingens knows, might be, for all he knows'; and so on. The individual concept in question is therefore potentially infinitely complex, and accordingly runs the risk of being finitely unlearnable.[16]

4. Conclusion

It is time to take stock. As we have seen, Lewis (1979) argued that we should take the contents of the attitudes to be, in effect, sets of centered worlds in order to account for the phenomena of self-location to which Perry (1977, 1979) drew attention. Stalnaker (2011), however, has argued that this would make communication theoretically impossible, and has suggested an alternative role for centered worlds in modelling the attitudes (2008, 2014). In particular, he has developed what I have been calling *Stalnaker models*, in which there is a relation of doxastic or epistemic accessibility holding between centered worlds, rather than uncentered worlds as in the more standard Hintikka models; and through the addition of *Stalnaker's constraint* on this relation he has ensured that the contents of these attitudes are sets of uncentered worlds, as desired. However, Stalnaker does not provide a semantic interpretation in these models for a language capable of expressing attitude ascriptions; and it is an exploration of this task that has been undertaken here.

After briefly describing the background to the discussion in the first section of the paper, in the second I sketched a Stalnaker model appropriate to capturing the relevant facts in a variant of Perry's (1977) case of Lingens the amnesiac lost in the Stanford library. I then provided a syntactic description of languages that can be used for making attitude ascriptions, and suggested a Millian semantics that predicted the intuitively correct truth-values for two key claims about the case. Yet this approach produced counterexamples to the schema T, which is widely held to express the factivity of knowledge, as well as to the transparency principles 4 and 5 which Stalnaker (2006) has endorsed for belief. These results were taken to be not only *prima facie* problematic for Stalnaker, but also just plain puzzling.

[16] It might be said that we should not be interpreting names and pronouns as expressing such non-rigid individual concepts. Instead, we should be reporting interpersonal attitudes using definite descriptions (which express such concepts). Perhaps. But the danger will then be that certain facts about iterated attitudes are ineffable: we may have no descriptions in our language that allow us to express the individual concepts that are needed.

In the third and final section of the paper I explored how the principles might be vindicated. It emerged that Stalnaker's models suggest a descriptivist strategy that might be thought to secure the transparency principles in particular. But this approach faces some difficulties of its own: we found that communication might turn out to be impossible in practice, even if not in theory; and that (interesting) common attitudes are not semantically predicted by individual attitudes, requiring as they do certain modes of presentation that are shared by all members of the group in question. Finally, given the robustness of the assumption required to secure transparency, the prospects for the descriptivist approach once it is dropped were explored: it was found that the language risked being either unproductive or unlearnable.

Where does this leave the question with which we began, regarding the role of centered worlds in the characterization of the attitudes? We have seen that Stalnaker attempts to strike a middle line between the approaches of Hintikka (1962), on which they play no role, and Lewis (1979), on which they figure in the contents of the attitudes. But his approach can now be seen to be intermediate between them in another respect. In particular, Taschek (2010) has suggested that there are broadly three approaches to semantics in general, and the semantics of names in particular: the Fregean view, which distinguishes sense and reference; the traditional Russellian, descriptivist view; and the neo-Russellian – or as I have been calling it, Millian – direct reference view. And the same might be said for attitudinal content.

As Taschek (2010) describes the Fregean position, sense is to be understood in a manner that ties it tightly to inferential role: accordingly, two sentences may have the same possible worlds truth-conditions, but differ in sense, if they can validly serve as premises or conclusions in different inferences (as 'Superman flies' and 'Clark flies' do); and so, two sentences can differ in sense even if they have the same possible world truth-conditions and the same constituency structure. Now, Hintikka's approach to the attitudes is standardly developed in line with a Millian framework. But given Taschek's characterization, Lewisian (centered worlds) content can be thought of as Fregean: for instance, the contents of the beliefs of Lewis' two gods differ, on Lewis' view, despite having the same structure, and despite being true (or at least false) in the same (uncentered) worlds.[17] But Stalnaker's constraint rules out precisely this kind of Fregean view of content; and so, as we have seen, Stalnaker's models naturally suggest a traditional Russellian descriptivist approach to the contents of the attitudes. While I have raised some concerns regarding this approach, I do not consider them to be decisive; but I hope to have at least clarified some of the consequences of the strategy, and in this respect at least, served to situate it.

Appendix

This appendix contains proofs of technical claims made in the paper.

Counterexample to 5

Consider the following Stalnaker model:

W = {@, x}
S = {Lingens, Perry}
E = {<Lingens, @>, <Perry, @>, <Lingens, x>, <Perry, x>}
R = {<<Lingens, @>, <Lingens, @>>, <<Lingens, @>, <Perry, x>>, <<Perry, @>, <Perry, @>>, <<Lingens, x>, <Lingens, x>>, <<Perry, x>, <Lingens, @>>, <<Perry, x>, <Perry, x>>}

[17] Moreover, Lewis (1979: 526) claims that contents are assigned to brain states in order to capture their causal roles; and we might consider the role of one attitude in causing another to be something like its *inferential* role.

And let the interpretation function I be such that:
I('Lingens') = Lingens
I('hungry') = the function f such that f(@) = {Lingens} and f(x) = {Perry}

In this model there is a counterexample to the schema:

(5) $If\ S\ does\ not\ believe\ (know)\ \varphi, then\ S\ believes\ (knows)\ S\ does\ not\ believe\ (know)\ \varphi.$

In particular, (a) 'Lingens doesn't know Lingens is not hungry' is true relative to <L,@>, while (b) 'Lingens knows Lingens doesn't know Lingens is not hungry' is false relative to <L,@>.

To see this, we consider the claims (a) and (b) in turn.

(a): 'Lingens doesn't know Lingens is not hungry' is true at <L,@> iff 'Lingens knows Lingens is not hungry' is false at <L,@>; iff 'Lingens is not hungry' is false at some <c,w> accessible from <L,@>; iff 'Lingens is hungry' is true at some <c,w> accessible from <L,@>. But 'Lingens is hungry' *is* true at some such <c,w>, namely <L,@>. The reason is that 'Lingens' denotes Lingens relative to <L,@>, and Lingens is an element of {Lingens}, which is I('hungry')(@).

(b): 'Lingens knows Lingens doesn't know Lingens is not hungry' is false at <L,@> iff there's a <c,w> accessible from <L,@> such that 'Lingens doesn't know Lingens is not hungry' is false relative to <c,w>; iff there's such a <c,w> such that 'Lingens knows Lingens is not hungry' is true relative to <c,w>; iff 'Lingens knows Lingens is not hungry' is true relative to either (i) <L,@> or (ii) <P,x>, (since <L,@> and <P,x> are accessible from <L,@>).

(i): 'Lingens knows Lingens is not hungry' is not true relative to <L,@>. That was shown in connection with claim (a).

(ii): 'Lingens knows Lingens is not hungry' is true relative to <P,x> iff 'Lingens is not hungry' is true at <c,w>, for every <c,w> accessible from <*Lingens*, x> (since Lingens is the denotation of 'Lingens' relative to <P,x>); iff 'Lingens is not hungry' is true at <L,x> (because <L,x> is the only such <c,w>). But Lingens (which is the denotation of 'Lingens' at <L,x>) is not an element of {Perry} (which is the extension of I('hungry') at x); so 'Lingens is not hungry' *is* true at <L,x>. So 'Lingens knows Lingens is not hungry' is true <P,x>. And this suffices for the truth of claim (b).

Counterexample to 4

Here we employ the following three-world model:

W={@, x, y}
S = {Lingens, Perry}
E={<L,@>, <P,@>, <L,x>, <P,x>, <L,y>, <P,y>}
R={<<L,@>,<L,@>>, <<L,@>,<P,x>>, <<P,@>,<P,@>>, <<L,x>,<L,x>>, <<L,x>,<L,y>>, <<P,x>,<L,@>>, <<L,y>,<L,y>>, <<L,y>,<L,x>>, <<P,y>,<P,y>>}

The interpretation function I is such that:
I('Lingens') = Lingens
I('hungry') = the function f such that f(@) = {Lingens}, f(x) = {Lingens, Perry}, and f(y) = the empty set.

Then we get a counterexample to the schema:

(4) $If\ S\ believes\ (knows)\ \varphi, then\ S\ believes\ (knows)\ S\ believes\ (knows)\ \varphi.$

In particular, 'If Lingens knows Lingens is hungry, then Lingens knows Lingens knows Lingens is hungry' is false at <L,@>. For (i) 'Lingens knows Lingens is hungry' is true at <L,@>; yet (ii) 'Lingens knows Lingens knows Lingens is hungry' is false at <L,@>.

(i): 'Lingens knows Lingens is hungry' is true at <L,@> iff for all <c,w> such that <L,@> R <c,w>, 'Lingens is hungry' is true at <c,w>; iff 'Lingens is hungry' is true at both <L,@> and <P,x>; iff Lingens (= I('Lingens') is an element of {Lingens} (= I('hungry')(@)) and Lingens (= I('Lingens') is an element of {Lingens, Perry} (= I('hungry')(x)). Obviously these conditions hold.

(ii): 'Lingens knows Lingens knows Lingens is hungry' is true at <L,@> iff for all <c,w> such that <L,@> R <c,w>, 'Lingens knows Lingens is hungry' is true at <c,w> (because Lingens is the denotation of 'Lingens' relative to <L,@>); iff for all <c',w'> such that <L,w> R <c',w'>, 'Lingens is hungry' is true at <c',w'> (because Lingens is the denotation of 'Lingens' relative to <c,w>, for all c, w). But <L,@> R <P,x> and <L,x> R <L,y>, yet 'Lingens is hungry' is not true relative to <L,y>, since Lingens is the denotation of 'Lingens' relative to <L,y>, yet Lingens is not an element of the empty set, which is I('hungry')(y).

Counterexample to T

Take the original model (i.e. the one involved in the counterexample to 5). Then we get a counterexample to the schema:

(T) $If\ S\ knows\ \varphi, then\ \varphi.$

In particular, consider the sentence 'if Lingens knows i is hungry, then i is hungry' evaluated relative to <Perry,@>: (i) the antecedent is true, because 'i' picks out Lingens in the (linguistic) context in which it occurs; but (ii) the consequent is false, since 'i' there refers to Perry. More fully:

(i) 'Lingens knows i is hungry' is true at <P,@> iff for all <c,w> such that <*Lingens*,@> R <c,w>, 'i is hungry' is true at <c,w> (since Lingens is the denotation of 'Lingens' relative to <P,@>); iff 'i is hungry' is true at both <Lingens,@> and <Perry,x>; iff Lingen is an element of f(@) = {Lingens} and Perry is an element of f(x)={Perry}. And these conditions hold.

(ii) 'i is hungry' is true at <P,@> iff the denotation of 'i' relative to <P,@> is in f(@); iff Perry is an element of {Lingens}. This condition does not hold.

Validity Proofs

No counterexamples to the usual principles expressing the reflexivity, transitivity, and Euclideanality of the accessibility relation can arise when only the special term 'i' is used. We take each of T, 4, and 5 in turn.

(T) If 'i believes φ(i)' is true at <c,w> then 'φ(i)' must be true at all <c',w'> such that <c,w> R <c',w'>. By reflexivity, it follows that 'φ(i)' must be true at <c,w>. But then 'if i believes that φ(i), then φ(i)' is true at <c,w>, for any <c,w>. In short, T is valid, when 'i' is the only term used.

(4) We can prove that 'if i believes φ(i) then i believes i believes φ(i)' is true at <c,w>, for all c, w, as follows. Suppose that 'i believes i believes φ(i)' is false at <c,w>, for some c,w. Then 'φ(i)' is false at some <c',w'> such that <c,w> R <c',w'> (since c is the denotation of 'i' at <c,w>). But then, there's some <c'',w''> accessible from <c',w'> such that 'φ(i)' is false relative to <c'',w''>. By transitivity of R, <c'',w''> is already

accessible from <c,w>. But then 'i believes φ(i)' is false at <c,w>. So, if 'i believes φ(i)' is true at <c,w> so is 'i believes i believes φ(i)'; and so, 'if i believes φ(i) then i believes i believes φ(i)' is true at <c,w>, for all c, w.

(5) Similarly, we can prove that 'if i does not believe φ(i), then 'i believes i does not believe φ(i)' is true at <c,w> for all c,w, as follows. Suppose 'i does not believe φ(i)' is true at <c,w>. Then 'i believes φ(i)' is false at <c,w>, and there's some <c',w'> accessible from <c,w> such that 'φ(i)' is false at <c',w'>. Now suppose (for reductio) that 'i believes i does not believe φ(i)' is false at <c,w>. Then there's a <c'',w''> accessible from <c,w> such that 'i does not believe φ(i)' is false at <c'',w''>, and so relative to which 'i believes that φ(i)' is true. But then 'φ(i)' is true relative to <c''',w'''>, for all <c''',w'''> accessible from <c'',w''> - including, by the Euclideanality of R, <c'w'>. This contradicts the claim above that 'φ(i)' is false at <c',w'>. So 'i believes i does not believe φ(i)' is true at <c,w>. So 'if i does not believe φ(i), then i believes i does not believe φ(i)' is true at <c,w>, for all c,w.

References

Egan, A., (2007). Epistemic modals, relativism, and assertion. *Philosophical studies*, 133: 1-22.

Frege, G., (1918). The thought: a logical investigation. Trans. P. Geach, *Mind*, LXV (259): 289-311.

Kaplan, D., (1989). Demonstratives. In J. Almog, J. Perry, and H. Wettstein (eds.), *Themes from Kaplan*, OUP, Oxford, pp. 481-564.

Lewis, D., (1979). Attitudes de dicto and de se. *The philosophical review*, 88(4): 513-43.

Moss, S., (2012). Updating as communication. *Philosophy and phenomenological research*, LXXXV(2): 225-48.

Ninan, D., (2010). De se attitudes: ascription and communication. *Philosophy compass*, 5(7): 551-67.

Perry, J., (1977). Frege on demonstratives. *The philosophical review*, 86(4): 474-97.

-(1979). The problem of the essential indexical. *Nous*, 13(1): 3-21.

Russell, B., (1910). Knowledge by acquaintance and knowledge by description. *Proceedings of the Aristotelian Society*, 11(5): 108-28.

Shoemaker, S., (1968). Self-reference and self-awareness. *The journal of philosophy*, LXV(3): 555-567.

Stalnaker, R., (1999). *Context and content*, OUP, Oxford.

-(2006). On logics of belief and knowledge. *Philosophical studies*, 128(1): 169-99.

-(2008). *Our knowledge of the internal world*, OUP, Oxford.

-(2011). The essential contextual. In J. Brown and H. Cappelen (eds.), *Assertion: new philosophical essays*, OUP, Oxford, pp. 137-150.

-(2014). *Context*, OUP, Oxford.

Taschek, W., (2010). On sense and reference: a critical reception. In Ricketts and Potter (eds.), *The Cambridge companion to Frege*, CUP, Cambridge, pp. 293-341.