Review of Robert Stalnaker, *Our Knowledge of the Internal World*

Ofra Magidor

- Penultimate draft. -

[Please cite only final version: *Philosophical Review* 119 (2010), pp. 384-391.]

1. Introduction

This is an extremely rich book, which offers a novel picture of knowledge of our internal world and how it fits into a more general picture of ourselves as both bearers and attributers of epistemic properties. The view is described as anti-foundationalist, thoroughly externalist, and deeply contextualist. The book combines some very broad-ranging insights in the philosophy of mind of language, with detailed discussion of the many of the deepest puzzles that have concerned philosophers in these fields over the last decades including Sleeping Beauty, Lewis's Two Gods, Kripke's Pierre, and Jackson's Mary.

The book is no easy read. The discussion is often very dense, material that appears later in the book helps shed light on earlier parts, and a good familiarity with Stalnaker's previous work is almost essential. It is only by reading the book patiently and carefully that the reader can put the pieces together into a general picture. However, the patience does pay off, as the picture that emerges is both intriguing and radical, and suggests some genuinely fresh solutions to a wide range of extensively discussed puzzles. I don't know if Stalnaker's views will be broadly adopted, but I expect they will certainly generate a lot of fruitful and exciting discussion.

2. Knowledge and self-location

The reader who is looking for a one-line summary of Stalnaker's view of knowledge of the internal world will be hard-pressed to find it. Rather than outlining a general position, Stalnaker considers a range of difficult puzzle cases and his position emerges gradually through his discussion of these cases. I will follow suit and try to give a sense of the main line of thought by considering Stalnaker's discussion of the Sleeping Beauty puzzle in comparison with Jackson's Mary.

First, consider Sleeping Beauty. As is familiar, Sleeping Beauty is put to sleep on Sunday night after being told that she will be woken up either once or twice in the next two days, depending on the flip of a fair coin. If heads, she is woken up only once, on Monday and if tails, she is woken up on Monday and again on Tuesday, but only after being given a drug that ensures that she will have no memory of the Monday waking. The question is to what degree should Sleeping Beauty believe, upon being woken up on Monday, that the coin will (or did) land heads. There are good reasons to think that the answer should be one third.[1] Yet this answer conflicts with the compelling intuition that upon being woken up on Monday, Sleeping Beauty gains no new information, and so her degree of belief in heads should remain just as it was on Sunday night – namely one half.

Stalnaker's solution to the puzzle offers a promising golden mean: he argues that the answer to the puzzle should be one third, but maintains that this answer can be defended while retaining the principle that Sleeping Beauty should only change her degree of belief in light of new information. On his proposal, upon being woken up on Monday, Sleeping Beauty *does* gain new information in the strictest sense of the term: she is able to rule out (uncentered) possible worlds that she wasn't able to rule out before.

The discussion of the solution (61-62) is not easy to follow, but let me try and rephrase it in my own terms. Suppose that when Sleeping Beauty wakes up on Monday, she names the token thought she is having as she first wakes up 'George'.[2] Now there are initially four genuine possibilities to consider: George might occur on Monday or Tuesday and the coin might land on heads or tails. But when she is woken up on Monday, Sleeping Beauty does gain new information: she learns that George is a token thought she is having *while being awake*, and so given that the combination Tuesday + heads + awake is not possible, she is able rule out the possibility where George occurs on Tuesday and the coin lands on heads.[3]

One thing to note about this solution is that it relies on some controversial assumptions about the metaphysics of token thoughts, for example that it is metaphysically possible for George to occur at a different time than it in fact did.[4] A more troubling question concerning this solution is why this line of reasoning was not already available to Sleeping Beauty on Sunday

[1] See e.g. Elga (2000) and Dorr (2002).

[2] Stalnaker insists throughout the book that the puzzles concerning self-location have nothing in particular to do with indexicality (see e.g. 59 and 87). I think he is completely right about this, and I therefore present his solution by appeal to names rather than to indexicals such as 'today' or 'this thought'.

[3] If one is worried that such token thoughts can only occur while one is awake one can instead assume that what Mary learns is that George exists (on this view, the initial four possibilities are: exists on Monday + tails, exists on Tuesday + tails, exists on Monday + heads, and doesn't exist + heads).

[4] Stalnaker suggests in page 70, n.1 that he is not relying on this assumption, but I cannot see how his solution to Sleeping Beauty goes through without it.

night. One possible response is to say that on Sunday night, Sleeping Beauty was not in a position to pick out the token thought George, and thus was not able to consider the different possibilities which concern it. However, it is far from clear that this response succeeds: after all, even on Sunday night Sleeping Beauty could pick out George by making the stipulation 'Let the first token thought that I have upon waking up on Monday be called 'Tony'', and infer that Tony (which is identical to George) is a thought she will have while being awake. Stalnaker, as I understand him, would resist this move by appealing to his usual diagonalization strategy. There is nothing to stop Sleeping Beauty from coining a name 'Tony' as suggested, but if she then says to herself 'Tony will occur while I'm awake' the proposition she thereby expresses is the trivial diagonal proposition which is true in a world *w* just in case whichever token thought she has in *w* upon being woken up on Monday occurs while she is awake, rather than the singular proposition concerning the token thought Tony/George. The stipulated introduction of the name 'Tony' is not sufficient to allow Sleeping Beauty to have the relevant singular thought concerning George that she is able to have on Monday.

Why then does the stipulation associated with 'George' on Monday entitle Sleeping Beauty to such singular thoughts (rather than to similarly diagnonalized ones)? One might be tempted to think that the difference has to do with the notion of *acquaintance*: the naming stipulation for 'Tony' involves an indirect description, while the naming stipulation for 'George' is made while in direct presence or acquaintance with the token thought. However, Stalnaker's discussion of Jackson's Mary case indicates that he resists this line of thought.

Stalnaker presents an interesting variant of the original Mary puzzle (86). As in the original case, Mary grows up in a black-and-white room, and never observes the colors red or green. Before being exposed to any colorful object, Mary coins the names 'Ph-red' and 'Ph-green' for the kind of phenomenal experience that a normal observer, with similar physical characteristics to those of Mary, would have upon seeing, in normal lighting conditions, a red or green object respectively. Mary is then told that she will be subjected to the following experiment: she will be shown either a red or a green star, depending on a flip of a coin. Suppose that following the coin flip Mary is shown a red star. Mary decides to name the phenomenal character of the experience she just had 'Wow'.

Is Mary now in a position to know that Ph-red is Wow? Given some common assumptions on the metaphysics of color, the proposition that Ph-red is Wow is a necessary one. Moreover, the crucial property involved in this claim (Wow) is one that Mary is as directly acquainted with as one can hope. Yet as Stalnaker is thinking about it, it would be counterintuitive to say that Mary now knows that Ph-red is Wow: all she learns after being shown the star is that either Ph-red or Ph-green is Wow. Thus despite her direct acquaintance with Wow, Stalnaker recommends diagonalization in this case: Mary's thoughts in terms of Wow are assessed by considering Ph-red relative to worlds where Mary is shown a red star, and Ph-green relative to worlds where she is shown a green star. This is precisely where Stalnaker's 'deep contexualism' about knowledge kicks in: there is no general recipe for which proposition 'that p' denotes in knowledge attributions of the form 'X knows that p'. The proposition in question depends on which aspects of the agent's knowledge and ignorance the knowledge-attributer is modeling in that particular context.

3. Externalism vs. Internalism

Stalnaker classifies his view as one that involves a "thorough going externalism" (111). Indeed, Stalnaker's discussion of the Mary coin-flipping scenario above reveals just how far his externalism goes. On Stalnaker's view, even after seeing the red star, there are some possible worlds compatible with Mary's knowledge in which the star she was shown was green. But because the connection between green and Ph-green is taken to by Stalnaker to be a necessary one, in those worlds Mary has an experience with green phenomenology. This in turns entails that even after seeing the red star, it is compatible with Mary's knowledge that her experience consisted of green phenomenology! Note just how extreme this is: Williamson's anti-luminosity arguments have already suggested that we are not always in a position to know precisely which phenomenal experience we are having.[5] But Williamson's arguments involve possibilities where the agent has a very similar phenomenal experience to the one they actually have. Stalnaker argues that Mary does not even know which of two radically different (red or green) phenomenal experiences she had.

It is surprising that despite this extreme externalism, there are also some highly internalist strands in Stalnaker's position. For one thing, his diagonalization strategy entails that at least in many contexts, words behave as if they had descriptive contents rather than referential ones – the kind of contents that are typically associated with internalism about content. More strikingly, in chapter 6, Stalnaker argues that it is a fundamental constraint on thought, that agents always know what they are thinking in the sense that their thoughts have the same content relative to each possible world compatible with their knowledge. But this assumption in turn requires something like the KK-principle (if an agent knows that p, then they know that they know that p). The reason is roughly this: in any case where the content of one's thought involves diagonalization, one diagonalizes over the set of worlds that are compatible with the one's knowledge. But if content is to be uniform across the worlds compatible with one's knowledge, the set over which one diagonalizes must be uniform too, and hence all worlds compatible with one's knowledge must agree with the actual world about which worlds are compatible with one's knowledge, and thus one must have knowledge that one

[5] Williamson (2001), chapter 4.

knows.[6] This is striking, because the KK-principle has been taken by many to be the hallmark of internalism about knowledge. Rather than being purely externalist, Stalnaker's view strikes me as an unusual blend of both internalist and externalist ideas.

4. Contextualism: superficial, moderate, and extreme

Stalnaker classifies his view of knowledge as a kind of 'deep contextualism', which he contrasts with 'superficial contextualism' (102-105). An example of superficial contextualism is Lewis's view of knowledge: on Lewis's view there is a default class of possibilities that the agent is able to rule out when she is correctly characterized as knowing that p. Of course in some contexts we allow that she can only rule out a subclass of those possibilities (those that she cannot properly ignore), but this contextualism is superficial in the sense that there is always one absolute, default class of possibilities one is considering (the class of possible worlds in which p is false). By contrast, Stalnaker's contextualism is one where "we need context, not to explain how we can go beyond our experience to eliminate possibilities, but to provide an account of the information that does the eliminating" (105). We have already seen an example of this above: Sleeping Beauty was able to eliminate possibilities via the singular proposition concerning the token thought George, while Mary, was only able to eliminate possibilities via descriptive material of the sort 'the phenomenal character of my experience – whatever it is'.

It seems to me, though, that there are two different ways to understand what Stalnaker means by 'deep contextualism'. According to one (call it '*moderate contextualism*'), for any agent and any time, there is one absolute class of possible worlds that the agent at that the time is not able to rule-out. On this view, the difficulty is merely in characterizing the relevant class of worlds. Both knowledge attributers and the agents themselves use language to characterize what agents know or learn in different situations. Thus, the suggestion goes, the claims 'X knows that p' (or 'X learnt that p') allow the class of worlds that the phrase 'that p' denotes to vary radically according to context.[7] The other position (call it '*extreme contextualism*'), lets contextualism run even deeper than this. On this view, there is no absolute, context-independent class of possible worlds that correctly characterizes each agent's total doxastic or

[6] See Hawthorne and Magidor (2009) for a much more detailed version of this argument.

[7] Of course, there is a sense in which this claim will be true for any case where 'p' contains an indexical or context-sensitive term. But the idea is that 'that p' can vary in unusual ways, ones that do not involve ordinary indexicality, and which depend on 'that p' being embedded in a belief or knowledge attribution.

epistemic state at a time: God herself could not say for each possible world whether it ultimately is or isn't one that belongs to the relevant state of the agent.[8]

I am not entirely clear which of the two views Stalnaker is defending. Some of his remarks suggest it is moderate contextualism. For example, he argues that "in the interpretation of statements of the form 'x believes that phi', the 'that phi' will denote a set of (uncentered) possible worlds…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". It is hard to see how modeling synchronic or diachronic comparisons of belief-states would be in any way straightforward if the set of possible worlds representing an agent's belief-state at a time would depend on subtle facts concerning the attributor's context, in the way that extreme contextualism suggests.

On the other hand, putting together the general picture Stalnaker paints throughout the book, I suspect that he is in fact pulled towards the more extreme view. To take a simple example: suppose Jill witnesses Jack's pants go on fire, and exclaims 'His pants are on fire!'. If this is not a context where we are particularly concerned with Jill's knowledge or ignorance of Jack's identity, then by Stalnaker's lines we can characterize Jill in this case as acquiring knowledge of the singular proposition which asserts of Jack (the very person) that his pants is on fire.[9] However, if we were to assume moderate contextualism (coupled with Stalnaker's general model of knowledge and belief), that would entail that Jill also learned, *for every essential property* phi of Jack, that the person whose pants were on fire has phi. This is so, because any possible world in which the person whose pants were on fire does not have phi is a world where the person is not Jack, and hence the world has already been ruled out by Jill's learning the singular proposition above. The way out of this problem seems to be a rejection of moderate contextualism in favor of extreme contextualism: only the latter allows that Jill's epistemic-state can be said to include only worlds in which the person whose pants are on fire

[8] Again, there are subtleties here. Any ordinary contextualist about knowledge would agree that the set of worlds which are compatible with an agent's 'knowledge' varies according to context. But many would still assume that there is some underlying context-invariant set of worlds representing the agent's epistemic state, which 'knowledge'-ascriptions can draw upon. Similarly, standard context-sensitive threshold models for belief assume a context-invariant underlying doxastic state, namely the agent's distribution of degrees of belief. At any rate, I use the terms 'moderate contextualism' and 'extreme contextualism' to distinguish between two interpretations of Stalnaker's view. This distinction may not transfer straightforwardly to other frameworks.

[9] Cf. Stalnaker's comment about Pierre having singular thoughts about Kiev on 111.

is Jack relative to one context of attribution, while it consists of another set (one that includes worlds where it is a duplicate of Jack on fire) relative to others.[10]

If I am right in characterizing Stalnaker's view as one of extreme contextualism, the view is indeed quite radical. As Stalnaker acknowledges, "This essential contextualist feature of the account…gives rise to a general worry that the externalist shift may involve a retreat from robust realism" (135). Few would take the fact that Tony Blair can, but I cannot, utter the sentence 'I used to be the prime minister' truthfully to undermine robust realism about British government offices: it is clear that Tony Blair and I would be describing the same underlying reality, albeit using different words. But on Stalnaker's view which fundamental epistemic or doxastic states an agent can be said to possess varies radically, and in fairly unsystematic and unconstrained ways according to the interests and concerns of the attributors of these states. This may suggest (even if not entail) that there isn't ultimately an underlying attributor-independent reality concerning such states, one that different attributors are all describing, albeit using different words.

One way to nevertheless reconcile robust realism about mental states with Stalnaker's extreme contextualism is to drive a wedge between the *real* mental states, and our representation of such mental states in terms of contents. That is to say, one can think of contents as merely another layer of modeling or representation, and not a fully realist description of the agent's epistemic or doxastic states. Yet it is hard to see Stalnaker opting for this kind of move. For one thing, he rejects the Fregean strategies for dealing with the various puzzles at the outset, complaining that they blur the line between the representation and what is being represented. [11]

At the end of the book, Stalnaker assures us that "the essentially contextual account of knowledge and of our intentional relations to the things we think about can be reconciled with a realist interpretation of knowledge and thought, but it takes philosophical work to do so" (135-6). I am sure many, like me, look forward to seeing how this philosophical vision develops in the coming years.[12]

Ofra Magidor, *Balliol College and the University of Oxford*.

[10] Stalnaker's remarks on 121 are probably the closest he gets towards an explicit endorsement of extreme contextualism.

[11] See 27-33.

[12] Thanks to Cian Dorr, John Hawthrone, and Sarah Moss for helpful discussion of this material.

References

Elga, Adam. 2000. Self Locating Belief and the Sleeping Beauty Problem. *Analysis* 60: 143-7.

Dorr, Cian. 2002. Sleeping Beauty: In Defence of Elga. *Analysis* 62: 292-295.

Hawthorne, John and Magidor, Ofra. 2009. Assertion, Context, and Epistemic Accessibility. *Mind* 118: 377-397

Williamson, Timothy. 2001. *Knowledge and its Limits*. New York: Oxford University Press.