Erschienen in: Mind. Oxford University Press (OUP). 2015, 124(496), pp. 1347-1352
https://doi.org/10.1093/mind/fzv055

Self-Knowing Agents, by Lucy O'Brien. Oxford: Oxford University Press, 2007. Pp. x + 231. H/b £53.00.

Self-Knowing Agents is a philosophically rich, substantial, rewarding and original book. It shows us a philosopher at work, in the best possible sense: we see a philosopher engaged in deep thought, and we are invited to witness, and participate in, the process of hard but satisfying philosophical labour.

The idea which lies at the heart of the book is that 'the clue to ourselves, and our self-consciousness, must lie in our agency' (p. vii). More specifically, as O'Brien announces at the outset, the book argues that 'the fact that we are active creatures gives us a way of being conscious of ourselves, through participation in our own actions, which we simply fail to identify if we concentrate on our purely passive ways of knowing' (p. vii).

However, the strategy which O'Brien employs to support this core claim is somewhat indirect. O'Brien does not 'start with our active natures in the effort to give an account of the ways we have of referring to and knowing ourselves' (p. viii), as one might expect. Instead, she sets out to:

> present problems we face in giving a satisfactory account of self-reference and self-knowledge and to consider existing solutions. The shape of the available solutions serves to bring out the way in which we have assumed that our relations to ourselves must either be perceptual or theoretical. It is in the face of the inadequacies of either of these options that the positive suggestions emerge. (p. viii)

Accordingly, part one of the book, entitled 'First-Person Reference', begins with four chapters in which extant views on first-person reference are discussed, and only then, in chapters five and six, does O'Brien set out to develop her own positive account, the 'Agency Account' of first-person reference. Part two of the book, entitled 'Action and Self-Knowledge' contains five further chapters (Chs. 7 to 11), which can all be understood as contributing to a development of the Agency Account, but which are so very far-ranging — chapter eight, for example, discusses the title question 'What are Actions?' — that the Agency Account itself does not receive quite the amount of careful attention that this reader at least would have wished for. More generally, I tend to think that a somewhat bolder and more direct approach in presenting the positive claims of the book might have made those claims more accessible, and might have made the author's own account more convincing. As it is, a lot of very careful attention is paid both (in Pt. 1 of the book) to the other accounts of self-reference with which the Agency Account competes, as well as (in Pt. 2 of the book) to the wide range of other philosophical issues which ideally should be considered if one wants to understand the Agency Account fully, but this careful attention to those two broad contextual themes arguably is at the cost of the Agency Account itself.

Within the limited space of the present review I therefore suggest that we do focus on the Agency Account, the positive account at the core of *Self-Knowing Agents*. The Agency Account is based on the suggestion that 'there is a form of awareness had by creatures capable of controlling their actions, mental and physical, that is independent of any capacity of the creature to understand the term or concept "I", that is both non-conceptual and non-perceptual in nature' (p. 76). O'Brien calls the relevant kind of awareness 'agent's awareness', and suggests that agent's awareness can 'immediately

warrant … the self-ascription of the action that the creature is aware of in this way' (p. 76), thus grounding our ability to refer to ourselves, and to have self-knowledge.

But then, what exactly *is* 'agent's awareness', and why should we accept that we have it? In contrast with passive phenomena, so O'Brien points out, active phenomena are characterized by the 'features of origin and control' (p. 116). A subject who engages in some activity, be it a mental activity or a physical activity, is the origin of the relevant activity, and the relevant activity is under the subject's control. For example, I ask you what day of the week it is today. You consider the question, and you come to judge that today is Tuesday. You are the origin of the relevant judgement, and in forming this judgement you have some rational control over what you do. For example, you have control over how the various pieces of evidence that come to mind are weighted, and whether or not you ultimately endorse the claim that today is Tuesday. Exemplary cases like the present one give us reason to accept that, more generally, a subject's mental and physical activities are indeed characterized by the 'features of origin and control'. But then, so O'Brien suggests:

> having this kind of control over one's mental life provides a primitive, representationally independent kind of awareness of what one is doing, and … this awareness is agent's awareness. (p. 117)

Why should we accept this? Why should having rational control over one's mental life entail a primitive awareness of what one is doing? And how might this primitive awareness be characterized; that is, what exactly *is* 'agent's awareness'?

In the main passage addressing these issues (p. 117), O'Brien offers a train of thought which might be reconstructed and summarized as follows:

(1) In order for a subject to have rational control over one of her mental activities, it is necessary that the subject be able to assess the various possibilities for action that are available in the relevant situation.

(2) Furthermore, in order for a subject to have rational control over one of her mental activities, it is also necessary that the subject be able to act immediately on her assessment of the various possibilities for action that are available once she has made such an assessment.

(3) But then, in order for a situation to 'make sense' (p. 117) in which a subject assesses various possibilities that are available in an attempt to determine how to act immediately, it is necessary that the relevant possibilities to act are possibilities for the subject herself.

(4) Thus, 'the very idea of an action produced by an active assessment [i.e. by an assessment for immediate action] by an agent carries with it the idea of an assessment by the agent of actions for her' (p. 117).

(5) Hence, 'any action produced directly on the basis of an active assessment [i.e. of an assessment for immediate action] by an agent will be an action … which the agent is aware of as hers' (p. 117), and this kind of awareness is here called 'agent's awareness'.
(6) Thus, an agent who exerts rational control over her mental life has 'agent's awareness'.
(7) But then, we all can, and sometimes do, exert rational control over our mental lives.
(8) Thus, we all do have 'agent's awareness'.

This train of thought certainly does address the question why having rational control over one's mental life should entail 'agent's awareness', as well as the question how 'agent's awareness' might be characterized. However, what we are offered here might require some further discussion, and I will focus on two issues which strike me as particularly important.

The first query I have concerns claim (1). The condition formulated by claim (1) seems a cornerstone of the positive argument developed in the present book. However, *why* should we accept claim (1)? Quite on the contrary, so an opponent might hold, in order for a subject to have rational control over her mental activities it is *not* necessary that the subject herself be able to assess various possibilities for action that are available to her, as claim (1) has it. (The opponent might go further and say that it is not even necessary that there *be* various possible ways in which the subject could act.) Rather, so the opponent might say, all it takes in order for a subject to have rational control over a certain mental activity is for the subject to have *a good reason* to engage in the relevant mental activity, for her to *choose* to engage in it, and for that choice to play a role in bringing the relevant action about. While O'Brien seems to take claim (1) to be obviously true, the opponent's opposing view as sketched does not seem to be obviously false, so in order to render the above train of thought more convincing, it would be useful to offer a response to the opponent's present considerations, as well as an argument in support of claim (1).

The second issue I would like to raise here concerns claim (5). As far as I can see, claim (5) is meant to follow from claim (4). But it seems that there are two possible readings of claim (4), and while claim (4) follows from claims (1) to (3) in only one of its two readings, only in the other of its two readings does claim (4) entail claim (5). This in turn suggests that the argument as a whole might be invalid.

The phrase that is ambiguous in claim (4) is the phrase 'of an assessment by the agent of actions *for her*'. Claim (4) does seem to follow from claims (1) to (3) if this phrase is understood to refer to actions which an *observer* might describe as 'actions for the agent' (i.e. 'actions for her'); by contrast, claim (4) does not seem to follow from claims (1) to (3) if the relevant phrase is

understood to refer to the *agent's own* conception of those possible actions as, from the agent's own point of view, 'actions for myself', a conception which we, as onlookers, might in turn describe as 'actions for her'. However, only in this latter reading might claim (4) possibly entail claim (5). This in turn suggests that the argument is not valid, and that some further work would be necessary in order to show convincingly that having rational control over one's mental life does entail having agent's awareness.

Offering a successful argument in support of this claim would seem important, for the claim that we do have 'agent's awareness' is certainly controversial, and so an opponent might well insist that unless a successful argument has been offered, we can simply deny that any such 'primitive awareness' occurs, or indeed is possible, at all. Instead, so the opponent might hold, sometimes processes of control which are said to provide the subject with 'agent's awareness' are in fact sub-personal processes, and at other times, if relevant processes of control do involve any relevant awareness on the part of the subject, the subject's relevant awareness is not at all primitive but rather does presuppose relevant concepts of self-reference.

O'Brien responds to the present objection by offering us three variations of a particular case of a mental 'transition' (pp. 117f.). 'Suppose', so she writes, 'that a subject has the beliefs *P, P* > *Q,* ∼*Q* and is rationally guided to drop her belief that ∼*Q* and accept *Q*' (p. 117). We can, so O'Brien suggests, conceive of this transition in at least three different ways: firstly, the transition might depend on sub-personal processes only; secondly, the transition might depend on the subject's forming beliefs about her three relevant beliefs ('I believe that P', etc.) and her then working her way through the problem on that basis. Thirdly, the subject might simply have the relevant first-order beliefs (i.e. the beliefs P, P > Q, and ¬Q), assess her options, and ultimately drop the belief that ¬Q and accept Q. In this third case, so O'Brien holds, the transition *cannot* be fully accounted for with reference to sub-personal processes, while it also does *not* depend on the subject's ability to refer to herself and her possession of relevant concepts. Instead, so O'Brien suggests, we can best account for this third case if we accept that the agent does have 'agent's awareness'.

However, a hard-core reductionist might well insist that everything that happens in the third case *can* be fully accounted for with reference to relevant sub-personal processes, and O'Brien has not offered us any reason as to why we should reject this view. On the other hand, once we grant that there is something about the case of the third transition which could not possibly be captured in purely sub-personal terms, an opponent might well argue that this is so because the third transition does depend on the subject's ability to refer to herself. In the third case the transition certainly does seem 'possible without the subject either self-ascribing, or indeed having the capacity to self-ascribe, *her beliefs*' (p. 118; my emphasis), as O'Brien says; but in the third case as described, the subject does assess 'possibilities she is aware of as

possibilities' (p. 118) which relate to what *she herself* might do — and an opponent might hold that in order for the subject to be able to even *select* any of the relevant possibilities as possibilities which should be considered, she has to understand that she is trying to consider possible ways of how *she herself* might act, and that, in order for a subject to understand this, the subject will have to rely on her ability to refer to herself, and thus will have to rely on relevant concepts of self-reference. The possibility of the third case therefore does not seem to prove that we have some primitive, non-conceptual awareness of ourselves, and so it seems that O'Brien's discussion of the present set of exemplary cases does not offer any further support for the claim that we do have 'agent's awareness' either.

A successful argument in support of that claim would, I think, strengthen the case of the book quite considerably. Of course, one might respond that the phenomenon of 'agent's awareness' is a basic and 'primitive' phenomenon, and maybe we should grant that there are some very basic phenomena which we can only somehow 'point at', without being able to offer an argument which shows that they actually occur. However, with respect to the (alleged) phenomenon of 'agent's awareness' the situation is somewhat more complicated, in that it seems that the cases of (alleged) 'agent's awareness' that have been 'pointed at' — for example, the third case of a 'transition' which we have just discussed — can be fully accounted for *without* the assumption that we have any such primitive 'agent's awareness'. Thus, while I find the suggestion very appealing that we do have a primitive awareness of ourselves and what we are doing when exerting rational control over our mental lives, an awareness which is 'non-conceptual and non-perceptual in nature' (p. 76), more should be done to support this claim.

If it was possible to establish more firmly that we do have 'agent's awareness', one should then ask in a next step whether agent's awareness could possibly play the fundamental role in grounding our ability to self-refer which O'Brien suggests that it does play. And indeed, one might have worries on this count as well. However, a discussion of those worries will have to be left for another occasion.

For the time being, I would like to conclude by urging anybody interested in the topics at the heart of the present book to give this book their continued and careful attention. As I said at the outset, *Self-Knowing Agents* is a substantial, rewarding and original book, and it will richly repay the reader's active engagement with it.

Department of Philosophy DOROTHEA DEBUS
University of York
Heslington
York YO10 5DD
dorothea.debus@york.ac.uk